DOCUMENT RESUME

ED 162 662 | IR 006 667

AUTHOR	Weiss, David J., Ed.
TITLE	Proceedings of the 1977 Computerized Adaptive Testing Conference.
INSTITUTION	Minnesota Univ., Minneapolis. Dept. of Psychology.
SPONS AGENCY	Office of Naval Research, Washington, D.C. Psychological Sciences Div.
PUB DATE	Jul 78
CONTRACT	N00014-76-C-0243, NR150-382
NOTE	454p.
EDRS PRICE	MF-$0.83 HC-$24.77 Plus Postage.
DESCRIPTORS	Ability; *Achievement Tests; *Computer Oriented Programs; *Educational Testing; *Group Intelligence Testing; Individual Tests; *Instrumentation; Performance Tests; Simulation; Test Bias

ABSTRACT

The 27 papers in this collection (26 of which were presented at the conference) are organized according to the eight topical sessions: (1) Improving Ability Measurement Using Different Item Formats, (2) Alternative Models for Adaptive Testing, (3) Psychological and Subgroup Effects, (4) Performance Testing by Interactive Simulation, (5) Implementations of Adaptive Testing, (6) Achievement/Performance Testing Viewed as a Classification Problem, (7) Achievement Testing Viewed as a Trait Measurement Problem, and (8) Computer-Based Testing as an Alternative to Paper-and-Pencil Testing. The document also includes discussions of these papers by R. Darrell Bock, Frederic Lord, Nancy Cole, Ernst Rothkopf, Richard Ferguson, Hariharan Swaminathan, and David Weiss; as well as a panel discussion entitled "Future Directions for Computerized Adaptive Testing," with presentations by Frederic Lord, Mark Reckase, Fumiko Samejima, Vern Urry, and David Weiss. Abstracts are included for each paper. (RAO)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED162662

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

PROCEEDINGS OF THE 1977 COMPUTERIZED ADAPTIVE TESTING CONFERENCE

held at the

University of Minnesota, July 19-22, 1977

EDITED BY

DAVID J. WEISS

PSYCHOMETRIC METHODS PROGRAM
DEPARTMENT OF PSYCHOLOGY
UNIVERSITY OF MINNESOTA
JULY 1978

Prepared under contract No. N00014-76-C-0243, NR150-382
with the Personnel and Training Research Programs
Psychological Sciences Division,
Office of Naval Research

Approved for public release, distribution unlimited.
Reproduction in whole or in part is permitted for
any purpose of the United States Government

TM006667

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Proceedings of the 1977 Computerized Adaptive Testing Conference		5. TYPE OF REPORT & PERIOD COVERED Conference Proceedings Report, 19-22 July 1977.
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Edited by David J. Weiss		8. CONTRACT OR GRANT NUMBER(s) N00014-76-C-0243
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Psychology University of Minnesota Minneapolis, MN 55455		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS P.E.:61153N PROJ.:RR042-04 T.A.:RR042-04-01 W.U.:NR150-382
11. CONTROLLING OFFICE NAME AND ADDRESS Personnel and Training Research Programs Office of Naval Research Arlington, VA 22217		12. REPORT DATE July 1978
		13. NUMBER OF PAGES iv + 450
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited. Reproduction in whole or in part is permitted for any purpose of the United States Government.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

This Conference and the preparation of these Proceedings were jointly sponsored by the Office of Naval Research and the Air Force Office of Scientific Research.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

testing
ability testing
adaptive testing
sequential testing
computerized testing
branched testing
individualized testing
tailored testing
programmed testing
response-contingent testing
automated testing
test bias
test fairness
achievement testing
performance testing

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This report is the Proceedings of the 1977 Computerized Adaptive Testing Conference held July 19-22, 1977, at the University of Minnesota. These Proceedings include 27 papers (26 of which were presented at the Conference); discussions of these papers by Darrell Bock, Frederic Lord, Nancy Cole, Ernst Rothkopf, Richard Ferguson, and David Weiss; and a panel discussion entitled "Future Directions for Computerized Adaptive Testing," with presentations by Frederic Lord, Mark Reckase, Fumiko Samejima, Vern Urry, and David Weiss.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

The papers are organized into the following topical sessions:

1. Improving Ability Measurement Using Different Item Formats
2. Alternative Models for Adaptive Testing
3. Psychological and Subgroup Effects
4. Performance Testing by Interactive Simulation
5. Implementations of Adaptive Testing
6. Achievement/Performance Testing Viewed as a Classification Problem
7. Achievement Testing Viewed as a Trait Measurement Problem
8. Computer-Based Testing as an Alternative to Paper-and-Pencil Testing

Acknowledgements

The 1977 Computerized Adaptive Testing Conference was jointly sponsored by the Office of Naval Research and the Air Force Office of Scientific Research. The Conference was held at the Nolte Conference Center on the campus of the University of Minnesota, July 19-22, 1977. The important contributions of the following individuals to the Conference and the preparation of these proceedings is gratefully acknowledged.

Conference Planning Committee	Charles Davis Brian Waters David Weiss
Nolte Conference Center Staff	Diane Campbell Joe Kroll
Art Work	Joel Brown William Correll Paul Kamin Karin Linnerooth LaRaye Osborne Jean-Pierre Pouplard Claudia Tysdal
Typing	Gerri Balter Karin Linnerooth LaRaye Osborne Claudia Tysdal Lynne Weyenberg
Photography	John Martin
Technical Editing	Barbara Camm

Contents

Conference Introduction	Marshall J. Farr	1
Session 1: Improving Ability Measurement Using Different Item Formats		2
Abstracts		3
Computerized Administration of Free-Response Items	C. David Vale	4
Interactive Testing Using Novel Item Formats	Charles H. Cory	15
The Applications of Graded Response Models: The Promise of the Future	Fumiko Samejima	28
Discussion	R. Darrell Bock	38
Session 2: Alternative Models for Adaptive Testing		45
Abstracts		46
An Empirical Evaluation of Implied Orders as a Basis for Adaptive Testing	Norman Cliff, Robert Cudeck, and Douglas McCormick	47
A Multivariate Model Sampling Procedure and a Method of Multidimensional Tailored Testing	Vern W. Urry	62
A Model for Testing with Multidimensional Items	James B. Sympson	82
Discussion	Frederic M. Lord	99
Session 3: Psychological and Subgroup Effects		104
Abstracts		105
Effects of Knowledge of Results and Varying Proportion Correct on Ability Test Performance and Psychological Variables	J. Stephen Prestwood	106
Student Attitudes Toward Tailored Testing	Bill R. Koch and Wayne M. Patience	116
Reduction of Test Bias by Adaptive Testing	Steven M. Pine	128
Discussion	Nancy S. Cole	143
Session 4: Performance Testing by Interactive Simulation		146
Abstracts		147
The Use of Simulation in Performance Testing	Christine McGuire	148
Problems of Performance Measurement in Computer-Based Simulation	Bruce W. Knerr	157
The Decision Measurement System as a Means of Testing Performance by Simulation	Myron A. Robinson and C. Lee Walker	172
Discussion	Ernst Z. Rothkopf	189
Session 5: Implementations of Adaptive Testing		193
Abstracts		194
Computerized Adaptive Testing and Personnel Accessioning System Design	Mark A. Underwood	196
Implementation of a Model Adaptive Testing System at an Armed Forces Entrance and Examination Station	Malcolm James Ree	216

Contents (continued)

Computerized Adaptive Testing with a Military Population .... Steven Gorman 221

Implementation of Tailored Testing at the Civil Service Commission .... Richard H. McKillip 231

Operational Considerations in Implementing Tailored Testing .... Harold Segal 235

A Low-Cost Terminal Usable for Computerized Adaptive Testing .... J. P. Lamos and B. K. Waters 238

Session 6: Achievement/Performance Testing Viewed as a Classification Problem 247

Abstracts 248

Applications of Sequential Testing Procedures to Performance Testing .... Kenneth I. Epstein and Claramae S. Knerr . 249

Adaptive Branching in a Multi-Content Achievement Test .... Roger J. Pennell 271

Adaptive Testing Applied to Hierarchically Structured Objectives-Based Programs .... Ronald K. Hambleton and Daniel R. Eignor .. 290

Multi-Content Adaptive Measurement of Achievement .... David J. Weiss and Joel M. Brown 312

Discussion .... Richard L. Ferguson ... 331

Session 7: Achievement Testing Viewed as a Trait Measurement Problem 336

Abstracts 337

Applications of Latent Trait Theory to Criterion-Referenced Testing .... James R. McBride 338

Unifactor Latent Trait Models Applied to Multifactor Tests: Results and Implications .... Mark D. Reckase 354

A Comparison of Conventional and Computer-Based Adaptive Achievement Testing .... Isaac I. Bejar 373

Discussion .... Hariharan Swaminathan . 386

Session 8: Computer-Based Testing as an Alternative to Paper-and-Pencil Testing 392

Abstracts 393

Research on Computer-Based Perceptual Testing .. David R. Hunter 394

Administering Paper-and-Pencil Tests by Computer, or the Medium is not Always the Message .... Jane Sachar and J. D. Fletcher 403

Discussion .... David J. Weiss 420

Panel Discussion: Future Directions for Computerized Adaptive Testing ... 423

Introduction 424

Frederic M. Lord 426

Mark D. Reckase 428

Fumiko Samejima 430

Vern W. Urry 441

David J. Weiss 442

Addresses of Conference Registrants 444

CONFERENCE INTRODUCTION

MARSHALL J. FARR, DIRECTOR
PERSONNEL AND TRAINING
RESEARCH PROGRAMS
OFFICE OF NAVAL RESEARCH

This conference, the second large-scale national meeting on computerized adaptive testing (CAT), is sponsored by the Office of Naval Research (ONR) and the Air Force Office of Scientific Research. The first CAT conference, held in June 1975, in Washington, D.C., was co-sponsored by ONR and the U.S. Civil Service Commission.

A number of participants at this Conference are ONR contractors, part of a programmatic thrust which has been spearheaded in the CAT realm. In addition, ONR has supported some of the key pioneering basic research that has laid the theoretical groundwork for CAT, supporting much of the test theory research from 1952 which led to Lord and Novick's (1968) "bible," Statistical Theories of Mental Test Scores, and supporting subsequent research by Lord on flexilevel testing.

I note with pleasure that attendance here is diverse and far-ranging; there are representatives from all of the military services, the Civil Service Commission, other federal civilian agencies, the university community, research and development commercial firms, and even from other governments.

SESSION 1

IMPROVING ABILITY MEASUREMENT USING DIFFERENT ITEM FORMATS

COMPUTERIZED ADMINISTRATION OF FREE-RESPONSE ITEMS

C. DAVID VALE
UNIVERSITY OF MINNESOTA

INTERACTIVE TESTING USING NOVEL ITEM FORMATS

CHARLES H. CORY
NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER

APPLICATION OF GRADED RESPONSE MODELS: THE PROMISE OF THE FUTURE

FUMIKO SAMEJIMA
UNIVERSITY OF TENNESSEE

DISCUSSION

R. DARRELL BOCK
UNIVERSITY OF CHICAGO

SESSION 1: ABSTRACTS

COMPUTERIZED ADMINISTRATION OF FREE-RESPONSE ITEMS

C. DAVID VALE

The theoretical advantages and practical problems encountered when administering test items in free-response rather than multiple-choice format are discussed. Comparisons are then made between tests in the two formats in terms of their psychometric characteristics. Conventional tests compared are a 20-item vocabulary test and a 130-item vocabulary test. A maximum-likelihood adaptive testing strategy for use with free-response items is developed and compared, in simulation, with a conventional vocabulary test. Future directions for free-response adaptive testing are suggested.

INTERACTIVE TESTING USING NOVEL ITEM FORMATS

CHARLES H. CORY

The research with computerized interactive measurement of abilities that has been conducted by the Navy Personnel Research and Development Center (NPRDC) is reviewed. Conclusions are drawn concerning the abilities for which computerized measurement has been most fruitful for supplementing paper-and-pencil tests. The objectives and methodology of a new follow-up study are discussed, and the characteristics of the new specially designed computerized testing equipment at NPRDC are explained and illustrated, as well as the rationale and characteristics of a new set of eight computer-administered tests.

THE APPLICATIONS OF GRADED RESPONSE MODELS: THE PROMISE OF THE FUTURE

FUMIKO SAMEJIMA

Using two weak parallel testing procedures, one of which uses a set of graded response items and the other a set of binary items, the effectiveness of the graded response item is examined in three different settings using the normal ogive model. The result shows the degree of effectiveness of the graded response items in comparison with the binary items with respect to such factors as the number of items required and the initial branching of examinees. Graded items were better than binary items in that they are less likely to fail in estimating an extremely high or low examinee ability level relative to the difficulty range of the item pool, a fact which is closely connected with the attenuation paradox of binary items.

COMPUTERIZED ADMINISTRATION OF FREE-RESPONSE ITEMS

C. DAVID VALE
UNIVERSITY OF MINNESOTA

The multiple-choice item format is extensively used in most large-scale tests of mental abilities. The primary advantage of this format is that it is easy to score. It has several disadvantages, however, including inability to effectively assess partial knowledge and vulnerability to random guessing. One approach which has been used in an attempt to remove the effects of guessing is formula scoring. A second approach (Waller, 1974) has attempted to eliminate the effects of guessing by simply not scoring those items on which the testee probably guessed, while a third has used confidence weighting and probabilistic response formats to assess partial knowledge in multiple-choice items. (See Bejar, 1975, for a review of these techniques.)

The present study is concerned with free-response items, the use of which renders success via random guessing virtually impossible and allows expression of partial knowledge. With the availability of a computer to administer and score items, simple formats such as multiple-choice are unnecessary; and testees can respond as freely as they would in an individualized oral examination. In free-response computerized testing, testees answer items in a free-response format by typing an answer on a computer terminal using their own language, rather than structured response alternatives.

The basic purpose of this study was to determine whether or not items administered in a free-response format could be scaled and scored in a manner efficient enough to make apparent their theoretical advantages over items administered in a multiple-choice format. The answer to this question was obviously dependent upon the nature of the item studied. An arithmetic item, having a single, unambiguous correct answer, can obviously be processed by computer as efficiently as a multiple-choice response alternative. Vocabulary items with several correct answers and the possiblity of misspellings are more difficult to process; and problem-solving items with semantic string responses are probably among the most difficult to process.

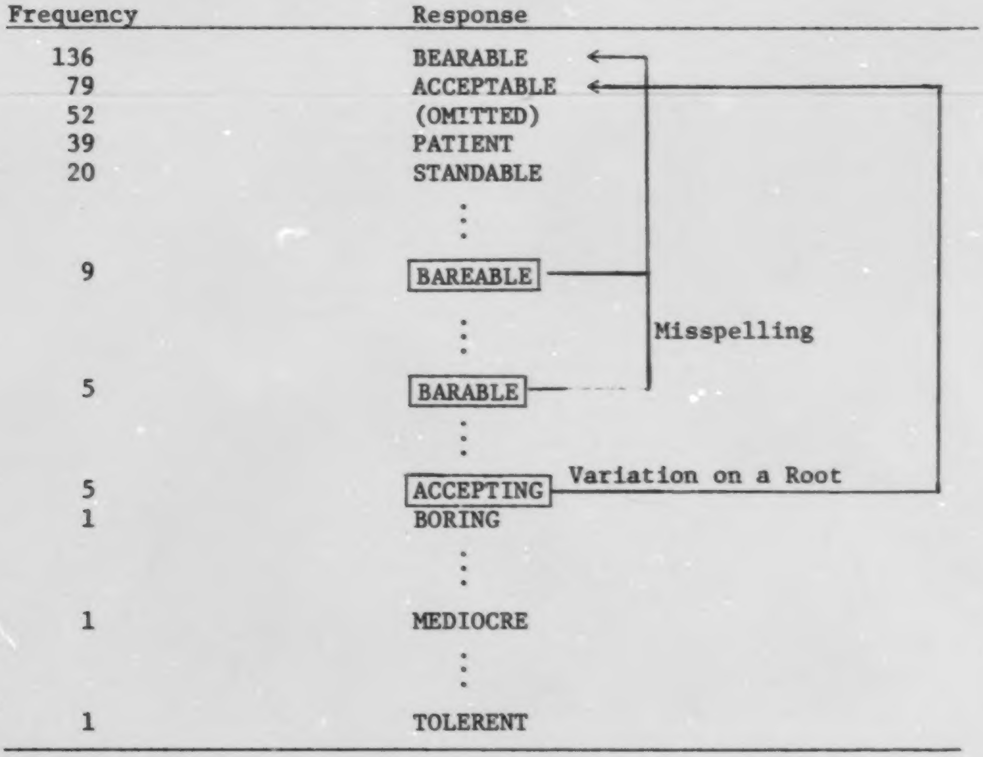
Study 1: Free-Response Vs. Multiple-Choice Vocabulary Items

Method

Test items. This research used free-response vocabulary items. (See Vale & Weiss, 1977, for details of this first phase.) Twenty multiple-choice items were selected and transformed into free-response items by removing their response alternatives. Testees were asked to respond to the items by typing into the computer a single word similar in meaning to the stem. These free-response items were then administered to 660 freshmen and sophomore college students. To provide a comparison, the 20 multiple-choice items from which the free-response items were made were also administered to the 660 testees. In order that the testees would not be able to use the multiple-choice alternatives when answering the free-response items, the free-response items were always administered first.

Reduction of responses. To each of the twenty free-response items an abundance of responses were given. Some of the responses which were given to one of the items are shown in Figure 1, ranked in order of frequency. As can be seen, the responses generated by the testees consisted of four types: (1) frequent responses, both correct and incorrect; (2) misspellings of frequent responses; (3) variations on roots of frequent responses; and (4) infrequent responses not included in Types 2 or 3. In order to reduce the number of responses to a manageable number of categories, responses were clustered on the basis of their formal similarity (i.e., the similarity between their strings of characters), based on a computer algorithm studied by Alberga (1967).

Figure 1
Most Frequent Responses to the Stem, Tolerable



Frequency	Response
136	BEARABLE
79	ACCEPTABLE
52	(OMITTED)
39	PATIENT
20	STANDABLE
	⋮
9	BAREABLE
	⋮
5	BARABLE
	⋮
5	ACCEPTING
1	BORING
	⋮
1	MEDIOCRE
	⋮
1	TOLERENT

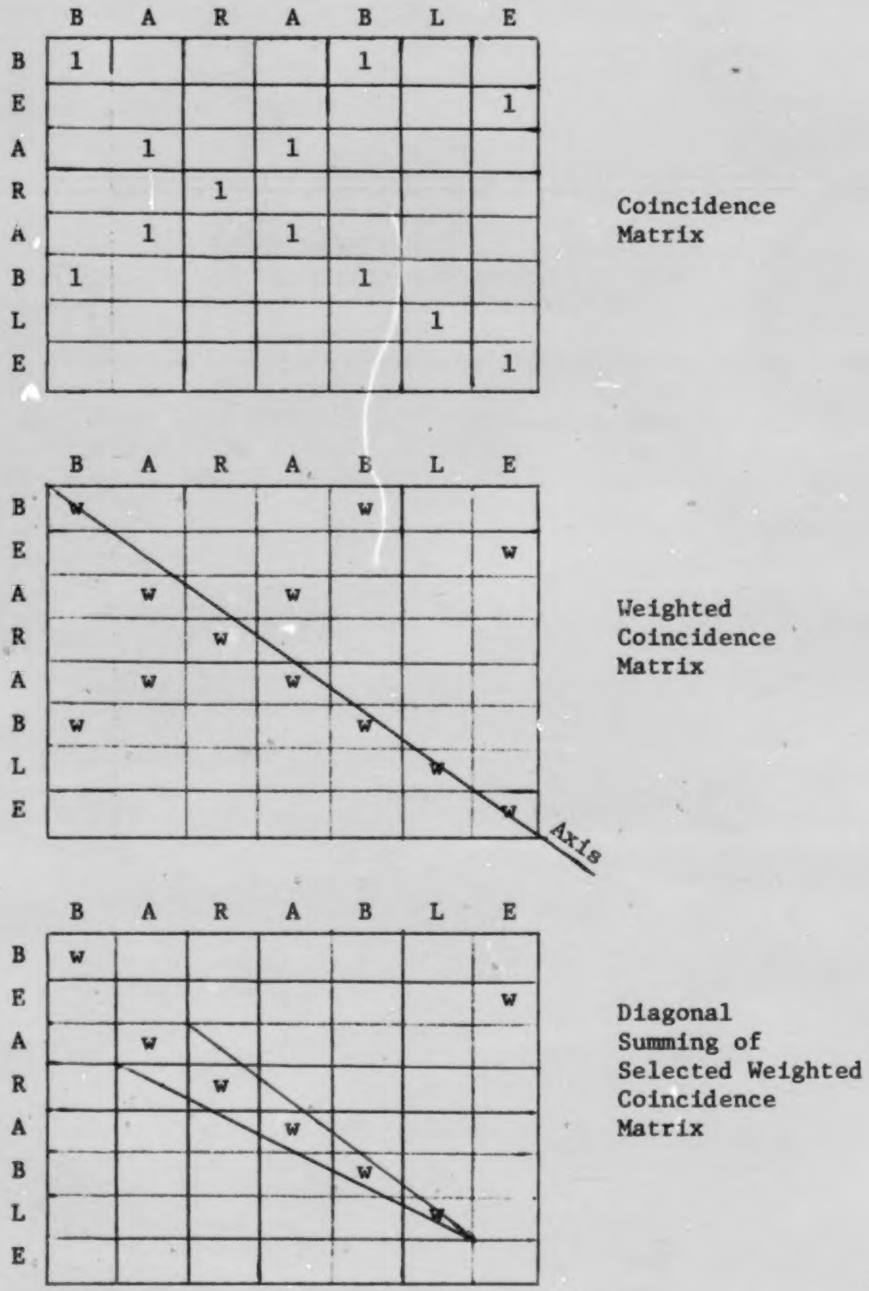
Alberga was looking for an algorithm to distinguish misspelling of a target word from a completely different word. He found or invented a list of 65 algorithms. Then, from an empirical data set of words and their misspellings, he evaluated the algorithms on their ability to recognize a misspelling of a target word without inappropriately labelling a different word as a misspelling of the target word. The algorithm which accomplished this the most successfully was Number 25.

In Algorithm 25 a coincidence matrix, such as that shown in Figure 2, is constructed with characters of the target word indexing the rows and characters of the test word, which in turn indexes the columns; the matrix need not be square. Elements of the matrix are 1 if the row and column characters correspond and 0 if they do not correspond. These elements are then weighted by their distance from the "axis" (the line from the upper left corner of the matrix to the lower right). This causes characters which occur in the proper position in the word to be given more weight. To eliminate the effects of repeated characters, the matrix is then processed in rows. The largest element in an, as yet, unchosen column is left unchanged; all others are set to 0. The similarity value is found by summing the remaining strings diagonally, weighting each element inversely to its position in its string, and normalizing the sum to constrain it to be between 0 and 1. From data reported by Alberga, discrimination is best when responses with similarity values greater than .12 are considered misspellings and those below .12 are considered unrelated words.

Using Algorithm 25 and beginning at the top of a list of responses ordered by frequency, responses were formally clustered. With the most frequent response of the cluster was grouped most of the misspellings of that response and most of the variations on its root. The eight most frequent responses and a catch-all category of "other" responses were then semantically clustered on the basis of semantic similarity into a maximum of six categories (attempting to keep the minimum category size at 30 testees). Words that were semantically very similar were clustered. Words infrequently used and definitely incorrect were clustered in the "omit" category; infrequent responses not semantically similar to any of the other categories and not definitely incorrect were clustered with the "other" category. Semantic clustering was performed by one person selected both for having the best available vocabulary and the time to devote to this task.

Scoring of free responses. To make these categories scorable, the items were calibrated according to Bock's (1972) polychotomous nominal response logistic model, using program LOGOG and an assumed normal distribution of ability. As a standard of comparison, the multiple-choice items were calibrated in the same way. The multiple-choice alternatives were too dissimilar to allow semantic clustering and thus were allowed to stand singly as a category (if 30 or more testees endorsed the response) or were grouped into both correct and incorrect categories. Calibration was again performed using LOGOG. The model underlying the calibration is a no-guessing model, which may not have been appropriate for multiple-choice items; it was used, however, to make the multiple-choice items comparable in terms of item information.

Figure 2
Matrices and Procedures of Algorithm 25



Coincidence Matrix

	B	A	R	A	B	L	E
B	1				1		
E							1
A		1		1			
R			1				
A		1		1			
B	1				1		
L						1	
E							1

Weighted Coincidence Matrix

	B	A	R	A	B	L	E
B	w				w		
E							w
A		w		w			
R			w				
A		w		w			
B	w				w		
L						w	
E							w

Diagonal Summing of Selected Weighted Coincidence Matrix

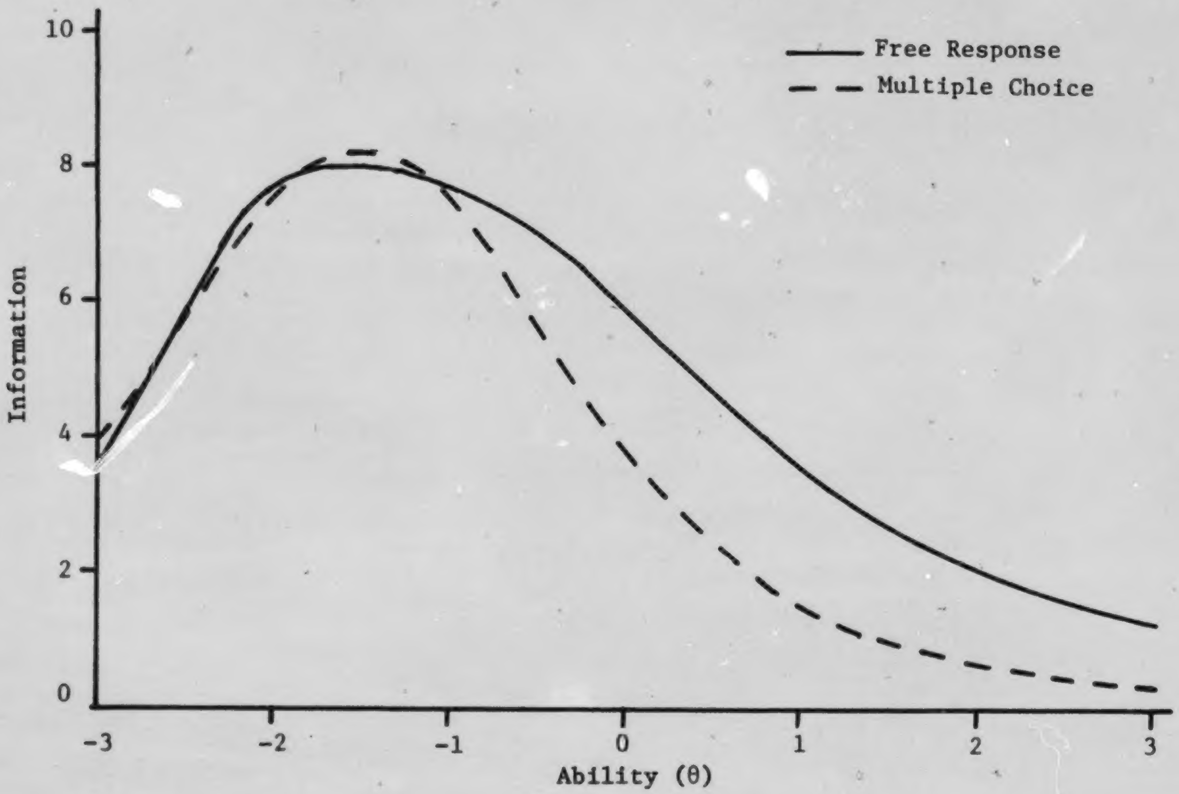
	B	A	R	A	B	L	E
B	w						
E							w
A		w					
R			w				
A				w			
B					w		
L						w	
E							

Results

Two free-response items and five multiple-choice items showed lack of fit with the model significant at $p< .05$ (see Vale & Weiss, 1977, pp. 6-7). Only one item was discarded, however, and then only because the fit was so poor that the parameters were unreasonably estimated. The remaining items were retained to avoid introducing a systematic bias into the comparison of the two item types.

Information. The first comparison between item types was in terms of the test information functions of the two 19-item tests. Item information functions were calculated for each item and were summed to obtain the test information functions. The resulting information functions are shown in Figure 3. Figure 3 shows that because both tests provided maximal information at $\theta=-1.5$, they were too easy for the population to which they were applied in this study. Assuming θ distributed normally with a mean of zero and standard deviation of one, the test information function should have peaked at $\theta=0$. More importantly, however, Figure 3 shows that while items administered in the two formats provided equivalent amounts of information near ability levels where the information functions peaked, the free-response items yielded more information at the higher ability levels.

Figure 3
Test Information Functions for Two 19-Item Tests



There are at least three possible reasons for the superiority of the free-response items: (1) they may have made better use of partial knowledge; (2) they may have eliminated error attributable to guessing; or (3) they may simply have fitted the model better because the no-guessing model was inappropriate for multiple-choice items. The first two reasons are legitimate superiorities; the last one is an artifact. The overall χ^2 statistics expressing lack of fit with the model (obtained from program LOGOG) were 651.40 with 560 degrees of freedom for the free-response test and 444.21 with 368 degrees of freedom for the multiple-choice test. *P*-values corresponding to these statistics were .020 and .013, respectively. This indicates that there was significant lack of fit for both tests, but that the fit for the multiple-choice items was only slightly worse than for the free-response items. Therefore, in all probability, the lack of fit did not account for the differences between the two information functions.

Reliability coefficients. To examine the data in the context of traditional test theory, split-half correlations were computed between two nine-item tests within each response format. Both maximum likelihood and number correct scores were used. For a free-response item, "correct" was defined as having provided an answer which was categorized into the response category determined a priori to be the "best" response category.

The results of the reliability analysis are summarized in Table 1. The highest correlation was between maximum-likelihood scores for the two free-response subtests. This value of .660 was significantly higher than the .598 obtained from maximum likelihood scoring of the multiple-choice test; however, it was not significantly higher than the .652 correlation which resulted from number-correct scoring of the multiple-choice test.

Table 1
Split-half Correlations Between Two Nine-Item Tests

	Free-Response	Multiple-Choice
Number correct	.424	.652
Maximum likelihood	.660	.598

That the number correct score of the multiple-choice items yielded a higher reliability than did the maximum-likelihood score could have been caused by (1) lack of fit with the model, (2) inaccurate estimation of the item parameters, or (3) both. Although 660 subjects may seem adequate for accurate estimation, it should be noted that some response categories for the multiple-choice test (each of which had two parameters to be estimated) were chosen by as few as 30 testees. For the free-response items, such small numbers of testees in a response category were even more prevalent.

The information comparisons are highly suggestive of the superiority of the free-response format. The reliability comparisons, however, suggest only a small difference between the two response formats.

Study 2: Replication

Method

Test items. To replicate these findings and to provide a pool of free-response vocabulary items suitable for adaptive testing, 130 items were administered to a group of testees similar to the group in the first study. The items came from a large pool of dichotomously calibrated multiple-choice items (McBride & Wiess, 1974) and were chosen to have discrimination (*a*) values between .4 and 2.8 and difficulty (*b*) values between -3.0 and 3.0. These items were administered with instructions similar to those of the previous study; however, the testees were given stronger instructions not to guess. Previously they had been told that, if they so chose, they could type a question mark rather than guess, but that the item would be counted incorrect. In this study, they were told to make an educated guess if they could, but not to guess if they had no idea which alternative was correct. This was done in an effort to improve the fit with the no-guessing model.

At the time of the analyses, only 130 testees had completed the items. This is insufficient data to provide accurate estimation of individual parameters and would make maximum-likelihood ability estimation look particularly poor. Estimation of parameters for these items should, however, give a reasonable idea of the distribution of item parameters and an acceptable estimate of test information functions.

Reduction of responses. The procedure for changing raw responses into item parameters was the same as in Study 1, except that the semantic clustering was performed on all responses given by more than one testee. Although this resulted in more work in the item construction process, it provided a more reasonable use of less frequent responses than the procedure used in Study 1 (in which only the eight most frequent categories were used). Visual inspection revealed that in some categories, some of the free-response items had too few responses to allow reasonable calibration with the number of testees available.

For the analyses reported here, the pool was narrowed to 100 items. Since program LOGOG allows a maximum of 50 items, the 100 items were divided into two sets. Four sets of items (two free-response and two multiple-choice) were calibrated in the same manner as in Study 1.

Table 2
Lack of Fit Statistics for Four Sets of Items

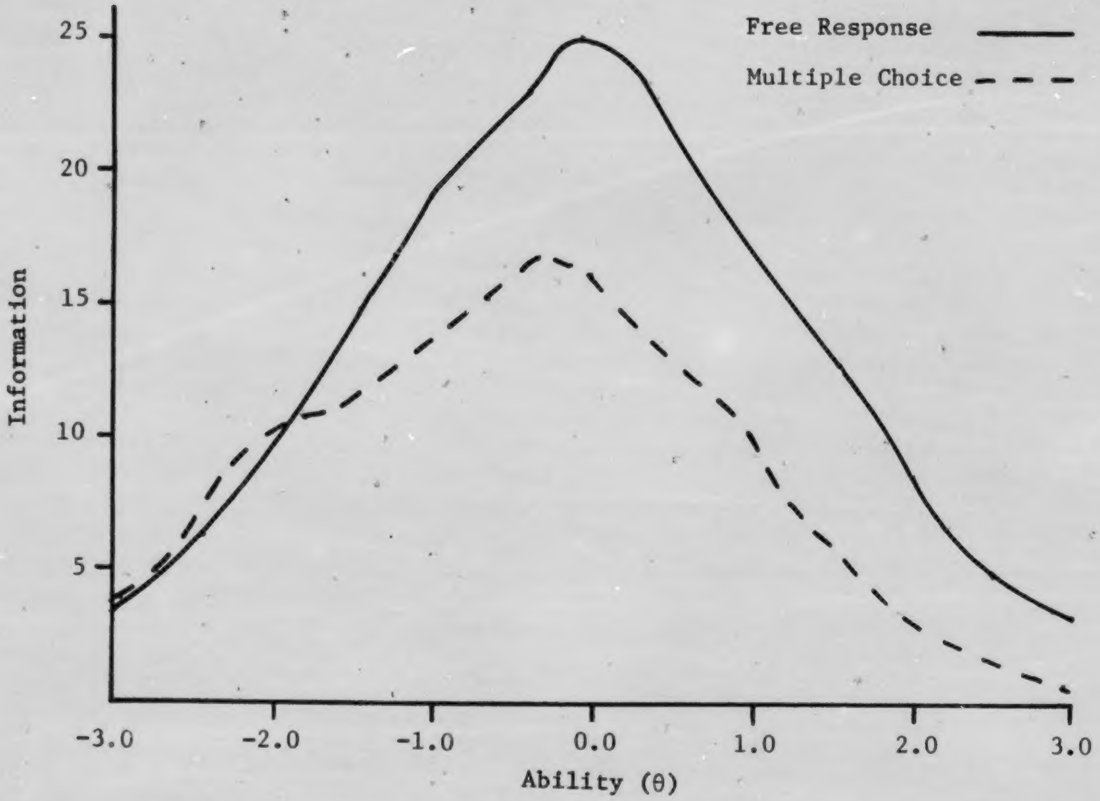
	Chi-Square	*DF*	*p*
Free-Response			
Set 1	1578.03	942	<.01
Set 2	908.35	950	.83
Multiple-Choice			
Set 1	452.70	398	.03
Set 2	401.54	398	.44

Results

Fit statistics for the four item sets are presented in Table 2. Set 1 showed significant lack of fit in both formats, with somewhat poorer fit in the free-response format. Set 2 did not exhibit significant lack of fit to the model. No single item had a profound lack of fit as had been observed in Study 1; consequently, all 100 items were retained for further analysis.

Information. Information functions for the two item types are shown in Figure 4; they are very similar to the information functions of Study 1. The primary difference is that the peaks were closer to θ=0.0, indicating that the test was more appropriate to the mean ability level of the group. Considering that the multiple-choice items fit the model essentially as well as the free-response items, these data suggest that the differences between the information functions were the result of the response mode, rather than an artifact of the lack of fit to the model.

Figure 4
Test Information Functions for Two 100-Item Tests



Reliability Coefficients. Reliability comparisons are presented in Table 3. The only meaningful conclusion to be drawn is that 130 testees were not enough to accurately calibrate polychotomous items. With these few testees, slightly more reliable scores were obtained in both response modes by summing the number of correct or "best" endorsements than were obtained by using maximum-likelihood scores.

Table 3
Split-half Correlations Between Two 50-Item Tests

	Free-Response	Multiple-Choice
Number Correct	.850	.915
Maximum Likelihood	.839	.903

Adaptive Testing with Free-Response Items

Dichotomous items with high information functions are constrained to have narrow information functions. No such constraint is imposed on polychotomous items: A polychotomous item can have an information function that is both high and wide. This suggests that when using free-response items, adaptive testing may provide less advantage over conventional testing. To evaluate this possibility, an adaptive testing strategy for free-response items was developed and compared to a conventional test by computer simulation. (See Vale & Weiss, 1975, for a discussion of computer simulations.)

Method

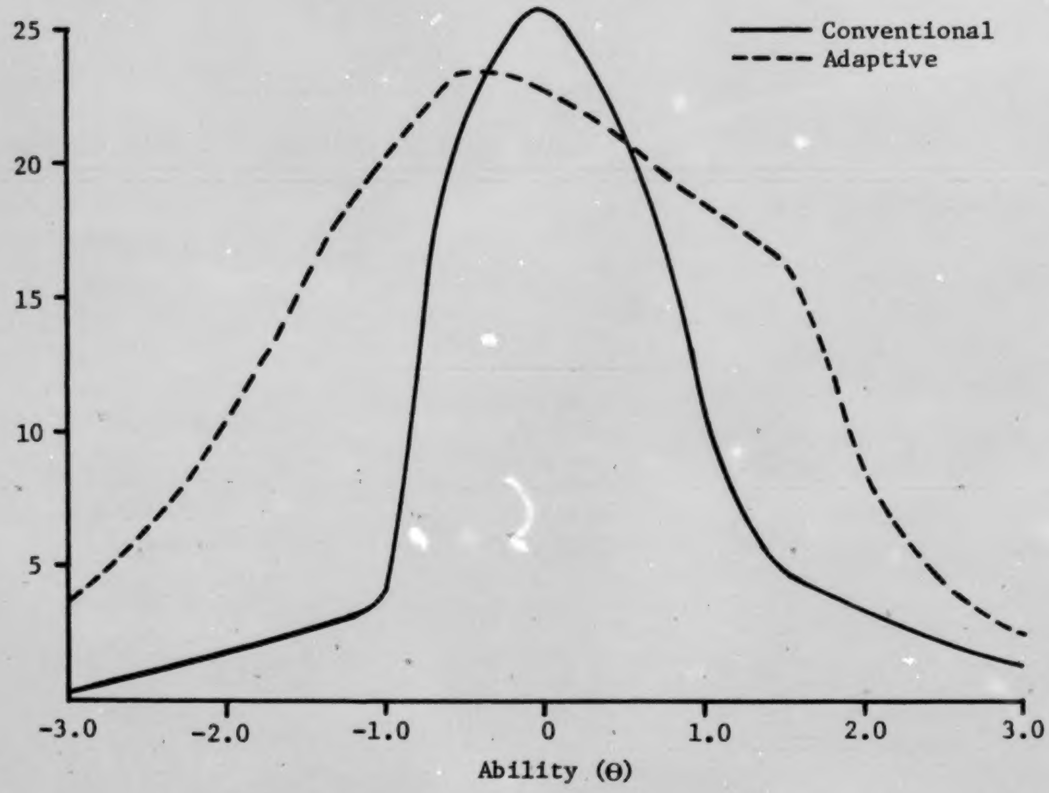
The adaptive strategy used was a maximum-likelihood strategy. In this particular strategy, the item providing the highest information at $\theta=0.0$ is administered first. Then a quasi-maximum-likelihood score (i.e., a maximum-likelihood score constrained to the range of -4.0 to 4.0) is calculated. The item providing the highest information at this estimate is administered next, and the procedure is repeated. In the simulation reported herein, the cycle was repeated until 20 items had been administered.

For comparison, a peaked conventional test was administered. This conventional test consisted of the 20 items from the pool of 100 that provided the highest level of information at $\theta=0.0$.

Results

The information functions for the adaptive and conventional tests are shown in Figure 5. As can be seen, the same conclusions can be drawn when using free-response items as when using multiple-choice items. A conventional test provides more information near where it is peaked (in this case between -.5 and .5), while the adaptive test provides more information elsewhere.

Figure 5
Information Functions for Free Response Tests
under Two Administration Strategies



Conclusions

Information analysis of free-response and multiple-choice items strongly suggests that free-response items are better than multiple-choice items. However, a large number of testees are needed to provide sufficiently accurate calibration in order to realize this advantage in terms of reliability; the data from 660 testees resulted in only slightly higher levels of reliability. Given adequate calibration, however, both free-response items and dichotomous multiple-choice items benefit from adaptive test administration.

References

Alberga, C. N. String similarity and misspellings. Communications of the ACM, 1967, 10, 302-313.

Bejar, I. I., An investigation of the dichotomous, graded, and continuous response level latent trait models. Unpublished doctoral dissertation, University of Minnesota, 1975.

Bock, R. D. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 1972, 37, 29-51.

McBride, J. R., & Weiss, D. J. A word knowledge item pool for adaptive ability measurement (Research Report 75-2). University of Minnesota, Department of Psychology, Psychometric Methods Program, 1975. (NTIS No. AD 781894)

Vale, C. D., & Weiss, D. J. A simulation study of stradaptive ability testing (Research Report 75-6). University of Minnesota, Department of Psychology, Psychometric Methods Program, 1975. (NTIS No. AD A020961)

Vale, C. D., & Weiss, D. J. A comparison of information functions of multiple-choice and free response vocabulary items (Research Report 77-2). University of Minnesota, Department of Psychology, Psychometric Methods Program, 1977.

Waller, M. J. Removing the effects of random guessing from latent trait ability estimates (Research Bulletin 74-32). Princeton, NJ: Educational Testing Service, 1974.

Acknowledgements

This study was supported by contract N00014-76-C-0243, NR150-382 from the Office of Naval Research, David J. Weiss, Principal Investigator, and by a grant of computer time from the University of Minnesota Computer Center.

INTERACTIVE TESTING USING NOVEL ITEM FORMATS

CHARLES H. CORY
NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER

Computerized equipment, when used for testing personnel, can provide a number of stimulus types and measurements that are not available from paper-and-pencil tests. For instance, computer-administered tests can (1) use a great variety of visual stimuli (including moving pictures), (2) interact with the test-taker to modify the presentation based on the testee's answers, and (3) provide detailed measurement of processing speeds at both test and item levels. Since these measurement characteristics correspond more closely with the task requirements of some jobs than do those available from paper-and-pencil tests, computerized tests should be more predictive than paper-and-pencil tests for certain types of job performance. The major focus of a research project at the Navy Personnel Research and Development Center (NPRDC) has been to investigate the usefulness of the unique measurement capabilities of computerized equipment for improving personnel selection and assignment decisions.

Previous Research

The initial study (Cory, 1977; Cory, Rimland,& Bryson, 1977) compared the computer-administered tests with paper-and-pencil tests in measuring five personal attributes, as defined by Mecham and McCormick (1969). These attributes (or abilities) were Short Term Memory, Closure, Perceptual Speed, Detection of Movement, and Dealing with Concepts/Information. The computerized tests that were developed to measure these abilities were designated as the Graphic and Interactive Processing battery (GRIP). They were used to predict on-the-job performance at both the job element and global criterion levels for enlisted personnel in the Electrician's Mate, Personnelman, and Sonar Technician ratings.

The results of these studies showed that the GRIP battery predicted performance of Sonar Technicians--at both job element and global criterion levels--significantly better than did either the paper-and-pencil classification tests that were being used operationally or the paper-and-pencil experimental tests that were selected to closely approximate the computerized tests. The GRIP battery, however, provided little or no improvement in predictability for Personnelmen and for Electrician's Mates. At present, NPRDC is conducting a follow-up research effort in an attempt to confirm the previous significant findings with respect to Sonar Technicians and to extend them to Operations Specialist, a closely related job. This new study has commenced after a delay of several years. The delay occurred because the

IBM 1500 system used for the previous research was no longer available, and suitable replacement equipment was not readily forthcoming. However, recently, a sophisticated computer-controlled multimedia system (CM^2), which had been developed at NPRDC to serve as a research test bed for advanced instructional techniques was modified to permit administration of computerized tests.

Current Computer System

The CM^2 system has four carrels, each with two cathode-ray-tube (CRT) screens. Instructions, via black and white stimuli, are presented on the top CRT; and visual (color) stimuli of TV quality are presented on the bottom. A varied set of state-of-the-art video sources are available, including videodisc and videotape inputs; thus it is possible to present both still and moving pictures. Also available are capabilities for graphics and for audio input. Testees' answers are input by means of a modified typewriter keyboard with alphanumeric and a variety of special characters.

The CM^2 system is interfaced with a PDP 11/45 by means of an adaptation of the UNIX software system. Programming is carried out in "C." The system is presently housed in a trailer next to the testing hall at the Naval Training Center in San Diego. In this location it is expandable by four to six carrels. Because of differences in the characteristics of the CM^2 and former IBM 1500 systems, the programs for all tests from the former GRIP battery, which were adapted for the present research, have been rewritten; and a trial administration of the battery is underway, to be followed by administration of tests for data collection. It is anticipated that the data collection phases of the research will extend over a 4-month period, with on-the-job follow-up being carried out about 10 months later. The new battery is being designated as GRIP-2, while the previous battery will be referred to as GRIP-1.

Characteristics of Sonar Technician Jobs

A review of the duties performed by Sonar Technicians (ST) suggests that it is not surprising that the GRIP-1 tests were so successful in predicting ST job performance. According to the official Navy job description, the ST's major duties are to operate, control, evaluate, and interpret the output of sonar and oceanographic equipment. This involves the detection, identification, and tracking of vessels that come within range of a sonar platform. On nearly all of the equipment in current use in the Navy, the sonar scans are presented as images on CRT screens. These images are based on sweeps, which are updated periodically. The ST's tasks are (1) to observe and interpret the images in order to detect targets and (2) to manipulate equipment settings in order to follow and identify the targets. The duties of the Operations Specialist (OS), which is the Navy's term for Radarman, are similar to those of Sonar Technician; but they involve considerably more clerical and administrative detail.

Researchers involved in the GRIP-1 research were interested in the supervisors' perceptions of the major elements of the ST rating. Table 1 presents the job elements that most supervisors indicated were important

Table 1
Job Elements Which Were Found in the
GRIP-1 Research to be Important for STs

Job Element	Number of Jobs For Which Important (Total N=37)
Attention to Details, Completing Work	36
Using Spoken Verbal Communication	36
Using Visual Displays	35
Using Pictorial Materials	32
Using Written Materials	30
Keeping Up-to-Date	30
Unusually Good Precision	29
Adjusting Machines/Equipment	29

for the ST job which they were rating, arranged in descending order by frequency of selection. Since the total number of STs in the sample was 37, there was general agreement on the importance of most elements for ST jobs. It is clear that supervisors perceived the principal elements of an ST job to be the making of visual distinctions (as expressed in the Using Pictorial Materials and the Using Visual Displays job elements) and the accurate performance of exacting work (as expressed in the Unusually Good Precision, Attention to Details, and Keeping Up-to-Date job elements).

In addition, results of interviews with experienced personnel have indicated that the most important characteristic for STs is the ability to identify the changes from scan to scan that indicate target objects. In other words, from the random patterns formed from the noise in the system, they must be able to separate out consistent patterns of variation. This implies that STs must have good short-term memory for visual stimuli.

These observations support the validity data from the GRIP-1 research. Of the four GRIP-1 tests that were significantly valid for STs, three measured Short Term Memory. One of these tests concerned visual stimuli. The other valid GRIP-1 test measured ability to deal with concepts/information. In addition, Counting Numbers, a paper-and-pencil test of Perceptual Speed, was significantly valid for STs.

The GRIP-2 Battery

The eight tests in the GRIP-2 battery are shown in Table 2. The first four tests are specific to the GRIP-2 battery, and the last four are modifications of GRIP-1 tests that showed promise.

Vigilance Test

Vigilance is an ability which is required for a variety of watchstanding tasks in the Navy. It is a particularly important quality for STs because their workday includes a number of 30-minute watchstanding periods in front

Table 2
Tests in the GRIP-2 Battery

1. Vigilance
2. Memory Search
3. Identifying Concepts
4. Scan-Count
5. Memory for Numbers-2
6. Password-2
7. Memory for Patterns-2
8. Recognizing Objects-2

of a CRT display. Although findings from studies such as those by Holland (1958) and Mackworth (1950) have indicated that substantial individual differences exist among personnel in regard to their vigilance capabilities, none of the operational paper-and-pencil classification tests used by the Navy has been very successful in predicting vigilance.

Buckner, Harabedian, and McGrath (1960) completed the most definitive study of vigilance heretofore carried out for the Navy. In that study, individual variations in watchstanding ability were investigated by means of a task that involved the detection of a change in brightness in a one-inch-square screen of light. Detectible changes occurred at irregular intervals at an average rate of one every five minutes. A major finding of the study was that the percentage of targets correctly detected began to drop within the first five minutes of watchstanding and continued dropping at a relatively constant rate until asymptote was reached.

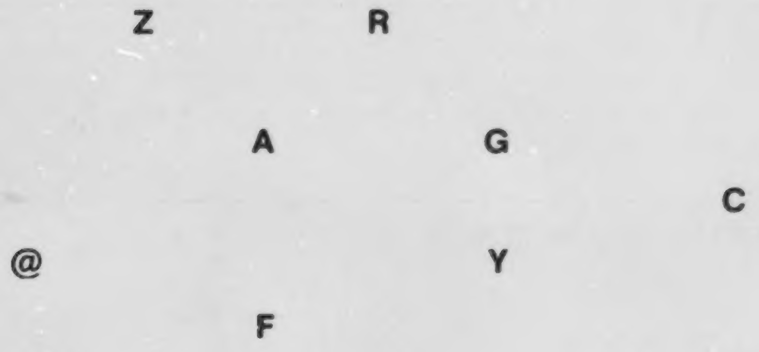
The computerized vigilance test is deceptively simple. An algorithm of the CM^2 equipment is used to generate capital letters at 1-second intervals and to distribute them randomly about the screen. At irregular intervals, averaging once every 2 minutes, an @ sign is presented instead of a letter. Two seconds later it is erased. The subject's task is to press the key as quickly as possible after seeing the @ sign. Figure 1 depicts the CRT display after 8 seconds have elapsed and immediately after the presentation of an @ character. To prevent overprinting characters, approximately every 30 seconds the screen is erased and the emission process begins again. The test continues for 20 minutes.

The test will be scored by computing percentages of hits and misses, number of false positives, and the average latencies for correct identifications.

Memory Search

Memory Search is an adaptation of a test developed by Sternberg (1966) in which several number series, each ranging from one to six digits, were presented. The digits in the presentation series were shown sequentially, one at a time, at 1.2 second intervals. Then after a time lapse, a probe series of digits was shown in the same manner. Some of the digits in the

Figure 1
CRT Display for Vigilance Test



probe series were in the original series and some were not. The subject's task was to indicate, as each digit in the probe series appeared, whether or not it was in the presentation series.

The Sternberg test had an error rate of only 1%, and its response latencies were almost entirely predicted by the length of the original span (1.2 seconds). In fact, the span length accounted for 99.4% of the variation of mean response latencies. Furthermore, positive responses, indicating matches between the digits in the presentation and the probe series, had average search latencies that were the same as those of negative responses. This is of interest because it indicates that the search process is exhaustive and continues through all 10 digits, even when a match has been found.

The Memory Search variable is of interest to the present research because (1) it appears to be a relatively pure measure of speed of information processing and (2) it is involved with Short Term Memory. Furthermore, because it requires sophisticated latency measurement, the test appears to be particularly appropriate for administration on computerized equipment.

Figure 2 shows items which are illustrative of the format of the Memory Search Test. The digits on the top line constitute the presentation series; and those on the bottom line, the probe series. As shown, three digits in the probe series also appear in the presentation series. This is characteristic of all probe series in the test. Although the number of digits in the presentation series varies from four to six, the number in the probe series is restricted to six, three of which also appear in the associated presentation series.

Sternberg's research was carried out within the tradition of experimental psychology, and he did not find substantial individual differences in performance. However, research by Hunt and Love (1972) has indicated individual differences to be somewhat greater than those found by Sternberg. Thus, to improve the Memory Search Test as a measure of individual differences the presentation spans have been set at three to six digits (versus one to six digits on the Sternberg test) and the length of exposure per digit has been reduced from 1.2 seconds to 1 second. Depending on a preliminary evaluation of results, it is also possible that the maximum presentation span may be increased to seven digits.

Figure 2
Illustrative Items for Memory Search

PRESENTATION SERIES	1 7 2 9 6 3
PROBE SERIES	4 5 1 8 7 3

Measures to be collected will be latencies of correct and incorrect answers, latency by length of digit span, average latency of response for each digit span, the total number of correct answers, and the number of correct answers for each length of digit span.

Identifying Concepts

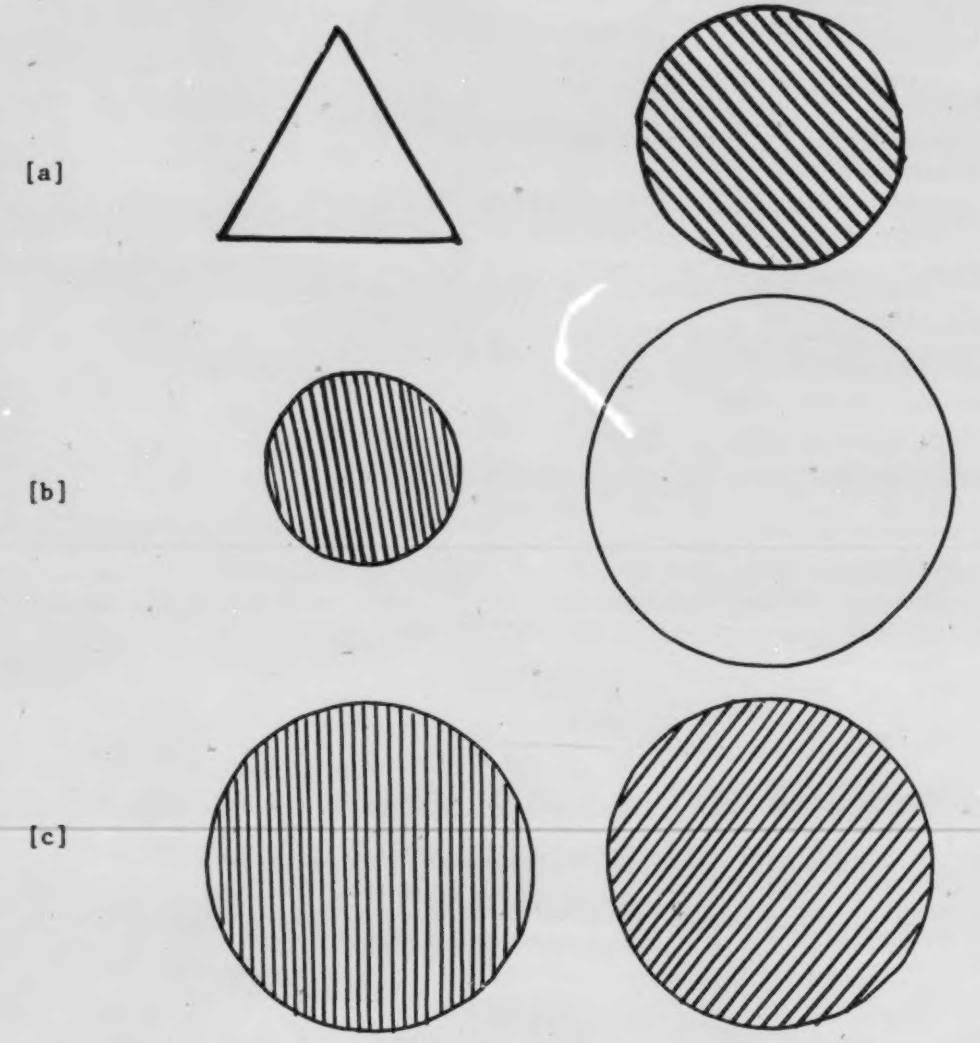
For the GRIP-1 research, the usefulness of computerized equipment for measuring the ability to deal with concepts/information was investigated by interactive tests that varied the presentation based on testees' responses. The tests were entitled "Twelve Questions" and "Password." A principal components analysis found that the two tests loaded on a separate component from other experimental and operational tests, thus supporting an interpretation that abilities required for interactive reasoning were somewhat different from those required for other reasoning tasks.

For the GRIP-2 battery it was desirable to replace the Twelve Questions Test with an interactive test that required less administration time and that had items with difficulty levels more appropriate to the average ability of the subjects. After an investigation of several alternatives, it was decided to adapt an experimental computerized test called "Concept Identification," which had been developed by the Air Force Human Resources Laboratory.

In the adapted test, which has been denominated "Identifying Concepts," a series of simple geometric shapes are presented, two at a time. The testee's task is to indicate whether or not the two are alike in terms of an unknown concept (e.g., both are squares, both have right angles). The testee is then given feedback concerning the correctness of the answer, and another pair of objects is presented. After the third pair, and each subsequent pair, the subject has an opportunity to type the name of the appropriate concept.

Figure 3 presents an example of the items in this test. In the first frame (Figure 3a), the two objects are about the same size but differ in shape and shading. Assume the testee makes the obvious guess that the two objects are different; the computer's response is "Right," and another two objects are presented (Figure 3b). In the second frame, the objects differ in size and shading but are the same in shape. Since the answer to the first item eliminates size as an answer, the question becomes whether or not the concept under consideration is shape (in which case the objects are the same) or shading (in which case they differ). Assume the testee then answers "Same," to which the computer replies "Wrong." The next frame (Figure 3c) would show objects that are the same in size, shape, and shading, but differ in slant. Should the answer be "Same" or "Different?" The answer to the previous frame should indicate that shape is not the concept under consideration. Assume that the testee answers "Same" to the third frame; he/she would receive a "Well done" from the computer and be invited to guess the name of the concept. At this point the testee should type "Shading."

Figure 3
Illustrative Frames from the Identifying Concepts Test



The presentation for a concept continues until 10 frames have been shown or until the concept has been identified, whichever occurs first. Scoring is in terms of the number of correct answers, the number of correct identifications, and the total latency of correct identifications.

Scan-Count

As has been mentioned previously, Counting Numbers, a paper-and-pencil test measuring perceptual speed, was found to be predictive for STs in the GRIP-1 research. Although scores on two other tests of perceptual speed evidenced substantial positive validities for STs, the validity of the Counting Numbers Test was the only one that was statistically significant. The relative-superiority of this test is believed to result from its inclusion of more searching and scanning tasks than the other two tests.

For the GRIP-2 research, it was decided to adapt Counting Numbers from a hand-scored to a computerized format in order to simplify the scoring and to permit study of performance latencies throughout the length of the test. Therefore, the computerized test Scan-Count was developed using the same type of items as in Counting Numbers.

Scan-Count (Figure 4) involves administration of a series of digit strings, each preceded by a target digit. The subject's task is to count the number of times the target digit occurs in the string. Thus, for the sample items shown, the correct answers would be 3 and 4.

Figure 4
Illustrative Items for Counting Numbers Test

1 13572563228435116

3 59047381473903235

Scoring will be based on the total number of correct answers, the total number of incorrect answers, the average latency for correct answers in each quarter of the test, the average latency for correct answers for the entire test, and the average latency for incorrect answers in each quarter of the test and in the total test.

Memory for Numbers-2

Memory for Numbers-2 is an adaptation of a similarly named test in the GRIP-1 battery. As in the original test, the adapted test consists of number spans having from 4 to 13 digits. The digits are presented at one-second intervals. Then, after a delay, the response frame is flashed on the screen. For example, as illustrated in Figure 5, the top line would be formed from four separate frames and the bottom line would be the response frame. The testee's task is to type the numbers in the same sequence as they appeared in the presentation span. Numbers typed by a testee appear over the dashed lines (cursors).

Figure 5
Illustrative Items for Memory for Numbers-2

PRESENTATION 2 5 7 3

RESPONSE — — — —

Previous research findings have indicated that a score based on the longest number span correctly recalled produces results that are comparable to those of other scoring methods, in particular to the more commonly used method of counting the total number of digits correctly recalled. Therefore, the Memory for Numbers-2 Test is scored by counting the longest digit span correctly recalled; and a one-up, one-down adaptive testing format is used to present the spans. Thus, for Memory for Numbers-2, the first presentation span has six digits. If the subject responds correctly, it is increased to seven digits. The test continues until the third presentation of a particular length of span has been made and responded to. Scoring consists of the last span correctly recalled, total latency for answers, and latencies for correct and incorrect answers.

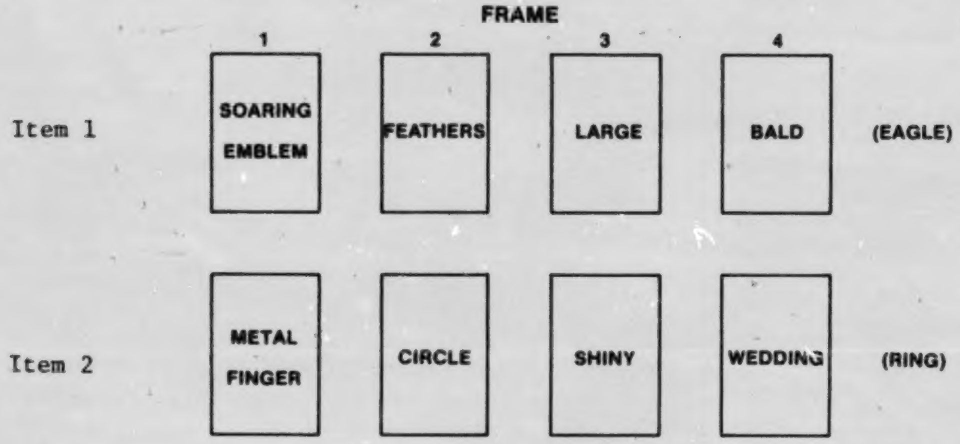
Password-2

Password-2 has the same format as Password-1, but it has a number of new items and several re-written ones. The tests resemble the "Password" game in that sets of words are shown which suggest a target word. The testee's task is to guess the target word by reasoning from the clues presented. Five clues are presented for each object. Figure 6 shows two sample items for Password-2 with the clues shown in order of presentation. As Figure 6 shows, the usual "Password" format is modified so that two clues are given initially instead of one. This change is intended to facilitate the convergent reasoning process.

For instance, for Item 1 the first clue, soaring, suggests that the object is a bird. The second clue, emblem, which identifies a use with which the bird is associated, should narrow the choice to a particular genus (eagle). Testees who do not decipher the implications of the clues in the first frame receive more specific indications of the physical characteristics of the bird in the second and third frames. If they still do not identify the object, the fourth frame suggests the best known species of the target word-- in this case, bald for eagle. Similar practices are followed in presenting Item 2.

Scoring consists of the total number of objects correctly identified, total number not identified, total clues required by the subject, total latencies for correct answers, average latency for correct answers, and total number of incorrect guesses.

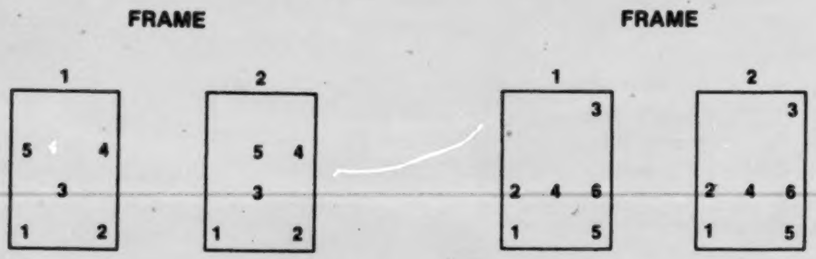
Figure 6
Illustrative Items for Password-2


FRAME
1
2
3
4
Item 1
SOARING
EMBLEM
FEATHERS
LARGE
BALD
(EAGLE)
Item 2
METAL
FINGER
CIRCLE
SHINY
WEDDING
(RING)

Memory for Patterns-2

The basic difference between Memory for Patterns-1 and Memory for Patterns-2 is that the presentation format of the latter does not require the subject to use a light pen, which had been available for the first test. In the revised presentation format, from four to six numbers are flashed sequentially, one at a time, to form a generally irregular, two-dimensional shape, as shown in the example (Figure 7). Then, a second pattern composed of the same numbers is shown in the same manner. All corresponding numbers in the two patterns may be located in the same place in their respective patterns, or one--and only one--of the numbers may be in a different location. The testee's task is to determine which, if any, of the numbers differ in location in the two patterns. In the first frame, then, the answer is "five" to indicate the difference in location. For the second item, the answer is "zero" to indicate that there are no differences between the patterns.

Figure 7
Illustrative Items for Memory for Patterns


FRAME
1
5
4
3
1
2
2
5
4
3
1
2
FRAME
1
3
2
4
6
1
5
2
3
2
4
6
1
5

Scoring will consist of the number of correct answers, the total latency of answers, and the average latency for both correct and incorrect answers.

Recognizing Objects-2

Although Recognizing Objects-1 was not as effective a predictor for STs as paper-and-pencil measures of Closure, the computerized test had validities that were high enough to suggest that a modified and improved Recognizing Objects Test could be useful for personnel selection. For Recognizing Objects-2, partially blotted-out pictures of common objects are presented. In the first presentation, 10% of the object is shown. In each successive frame, 5% more detail is added to the object. After each presentation frame, the subject has an opportunity to identify the object by typing its name.

For instance, Figure 8a shows a 10% view of a common item that is fairly difficult to recognize. When 20% of the detail is shown (Figure 8b), the object can be perceived somewhat more clearly, but still will not be identified by some people. However, when 35% (Figure 8c) of the detail is shown, most, if not all, testees can recognize a comb.

Scoring counts the total number of frames shown prior to identification of the objects, the total viewing time to initiation of typing the correct answers, the total number of wrong answers volunteered, the correct answers expressed as a proportion of total answers, and the average viewing time to correct answer.

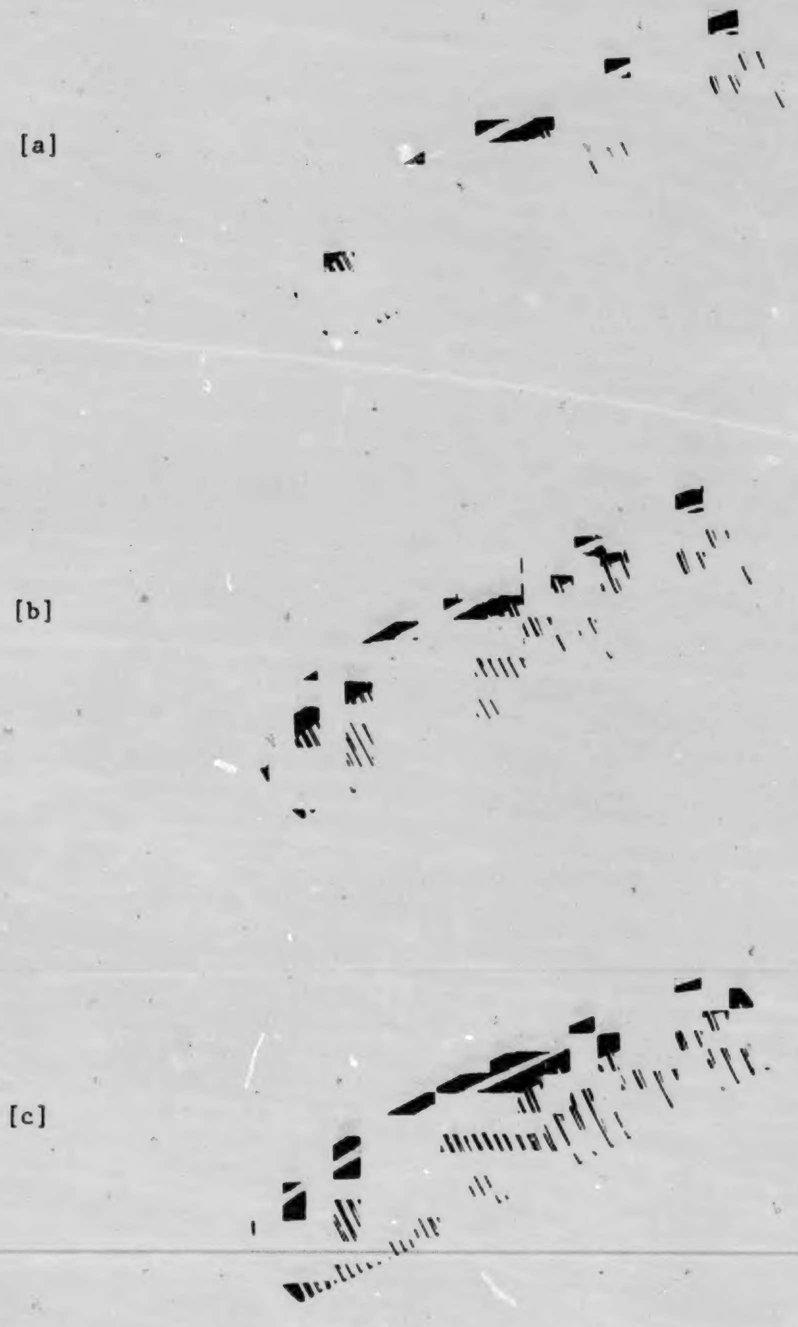
Discussion

To prevent typing speed from contaminating the computed speed of mental processes, latencies are timed from the original presentation of the stimuli to the onset of typing. Spelling requirements have been reduced to a minimum for all of the tests. All objects which require free answers have simple, easily spelled, one-word names having, for the most part, no more than five letters. The answer tables for these questions include not only the correct spelling, but also every misspelling that could be either conceived as plausible by the test technician or identified from an experimental administration of the tests.

Of necessity, heretofore it has not been possible to develop latent trait item characteristic curves for these types of items. Consequently, it has not been possible, for the most part, to present them by means of adaptive branching methodologies. However, a number of these item types may be appropriate for latent trait measurement and consequently for adaptive administration.

The total number of subjects to be tested for the present research is 400. If another study is carried out using a sample of similar size, it should be possible by pooling samples to compute stable item characteristic curves for many of the tests and to estimate the potential usefulness of adaptive branching methodologies for these non-traditional item types.

Figure 8
Illustrative Item from the Recognizing Objects-2 Test



References

Buckner, D. N., Harabedian, A., & McGrath, J. J. Human factor problems in anti-submarine warfare: A study of individual differences in vigilance performance (Technical Report 2). Los Angeles, CA: Human Factors Research, Inc., 1960.

Cory, C. H. Relative utility of computerized versus paper-and-pencil tests for predicting job performance. Applied Psychological Measurement, 1977, 1, 551-564.

Cory C. H., Rimland, B., & Bryson, R. A. Using computerized tests to measure new dimensions of abilities: An exploratory study. Applied Psychological Measurement, 1977, 1, 101-110.

Holland, J. G. Human Vigilance. Science, 1958, 128, 61-67.

Hunt, E., & Love, T. How good can memory be? In A. W. Melton & E. Martin (Eds.), Coding processes in human memory. Washington, DC: V. H. Winston & Sons, 1972.

Mackworth, N. H. Researches on the measurement of human performance. Medical Research Council Special Report, Series No. 268, 1950.

Mecham, R. C., & McCormick, E. J. The rated attribute requirements of job elements in the Position Analysis Questionnaire (Report No. 1). Purdue University, Occupational Research Center, January 1969. (NTIS No. AD 682-490)

Sternberg, S. High speed scanning in human memory. Science, 1966, 153, 652-654.

Acknowledgements

The opinions or assertations contained herein are those of the writer and are not to be construed as official or reflecting the views of the Navy Department.

THE APPLICATION OF GRADED RESPONSE MODELS:
THE PROMISE OF THE FUTURE

FUMIKO SAMEJIMA
UNIVERSITY OF TENNESSEE

Weak parallel tests have been introduced by Samejima (1977b) in contrast to strong parallel tests in the context of latent trait theory. Two tests are *strongly parallel* if (1) they have the same number of items and (2) a one-to-one correspondence of each item on the first test can be made with one and only one item on the second test with respect to the identity of the number of item score categories and the set of operating characteristics of item score categories. On the other hand, *weak parallel tests* are any pair of tests measuring the same ability or latent trait for which test information functions are identical. Thus two weak parallel tests can have (1) different numbers of items or (2) *no* one-to-one correspondence between the two sets of test items with respect to number of score categories or to sets of operating characteristics of the item scores.

It has been pointed out by Samejima (1977a) that in tailored testing, or computerized adaptive testing, any number of weak parallel tests can be made by prearranging a certain amount of test information and using it as the criterion in terminating the presentation of items to individual subjects. In such procedures two different item pools are not even needed; two item pools developed for measuring the same ability or latent trait will serve just as well.

Estimating Item-Operating Characteristics

Two methods of estimating operating characteristics have been presented by Samejima, using maximum likelihood estimation of ability or the latent trait and the constant test information function of a test pre-arranged for this purpose. One of the main objectives of these two methods is to estimate the operating characteristic efficiently without using a large number of subjects for its calibration. In both methods, a test is needed which would measure the ability in question and which would have a test information function that is substantially large and constant over the range of ability to be measured. In this situation, each individual's ability is estimated from his/her response pattern by maximum likelihood estimation. Thus

$$\hat{\theta} = \theta + \varepsilon \text{ ,} \qquad [1]$$

where θ is the individual parameter or the ability of the subject;
$\hat{\theta}$ is its maximum likelihood estimate; and
ε is the error of estimation.

The methods utilize such an asymptotic property of the maximum likelihood estimate that it distributes normally with θ and utilizes the inverse of the test information function for the two parameters (Samejima, 1975). The maximum likelihood estimates of the subjects are obtained. The subjects are then categorized into (m_g+1) groups in accordance with their performance on the new test item, g, for which the operating characteristics are to be estimated, where m_g is a positive integer and the highest score of item g.

In the first method (Samejima, 1977c), for each of (m_g+1) groups of maximum likelihood estimates, the normal approximation is made for the joint distribution of $\hat{\theta}$ and ε; and the error is calibrated accordingly by monte carlo methods to obtain the estimated frequency distribution of ability for each group. In the second method (Samejima, 1977d) the probability density function of $\hat{\theta}$ is graduated by a polynomial of degree 4 or 3 by the method of moments (Elderton & Johnson, 1969), in which, unlike the original method, the raw data is used instead of the frequency distribution. From this estimated probability density function, the first and second conditional moments of ε, given $\hat{\theta}$, are estimated by means of

$$E(\varepsilon|\hat{\theta}) = -\sigma^2\left\{\frac{d}{d\hat{\theta}}\,g(\hat{\theta})\right\}\left\{g(\hat{\theta})\right\}^{-1} \qquad [2]$$

and

$$E(\varepsilon^2|\hat{\theta}) = \sigma^4\left\{\frac{d^2}{d\hat{\theta}^2}\,g(\hat{\theta})\right\}\left\{g(\hat{\theta})\right\}^{-1} + \sigma^2, \qquad [3]$$

where $g(\hat{\theta})$ is the probability density function of $\hat{\theta}$ and
σ^2 is the inverse of the constant test information described earlier.

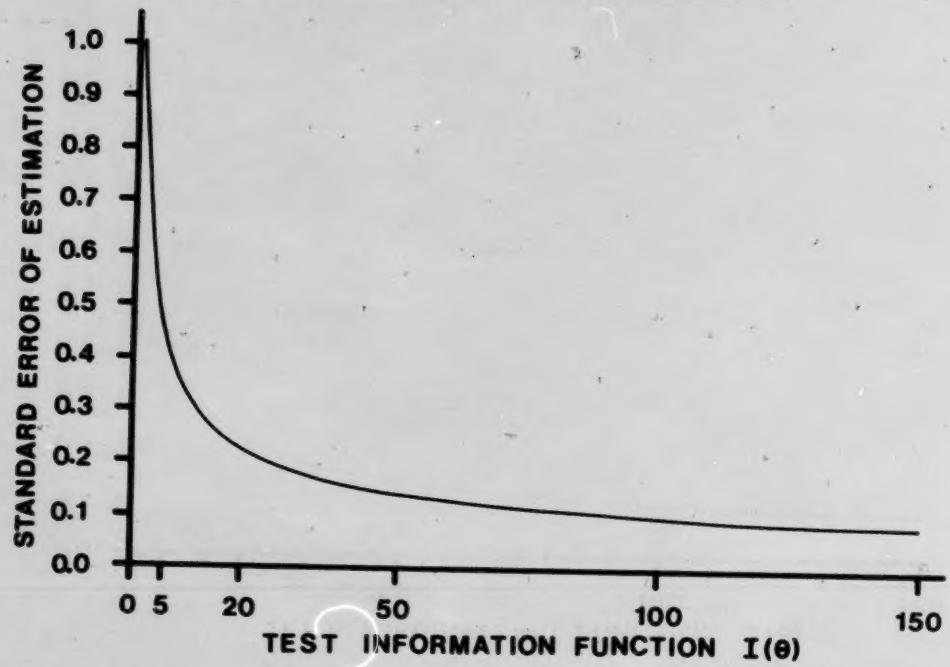
A Beta function is assumed for each conditional distribution of ε, given $\hat{\theta}$, with the lower bound $-2.55\sigma^2$ and the upper bound $2.55\sigma^2$. The other parameters are estimated from the above two conditional moments (Johnson & Kotz, 1970). The error is calibrated, following the Beta distribution thus estimated, to obtain the estimated frequency distribution of ability for each score group of a new item g. Some monte carlo studies have proved that both of these two methods work well, providing fairly accurate estimates of operating characteristics.

It should be noted that the above methods will be useful in tailored testing when there is a need to estimate the operating characteristics of new items to be added to an item pool. Since it cannot be expected that more than a few hundred subjects would be used for such a purpose in computerized adaptive testing, an efficient method of estimating the operating characteristics, such as that described, is all the more important.

Test Information Function and Standard Error of Estimation

One of the many advantages of latent trait theory over classical test theory is that the standard error of measurement is defined more meaningfully, as a function of the latent trait θ. It is defined as the square root of the inverse of the test information function. This measure is most meaningful when the test information function assumes a high enough value so that the conditional distribution of ε, given θ, is approximately normal (Samejima, 1977a, 1977c). When a prearranged value of the test information function is used as the criterion for terminating the presentation of new items in adaptive testing, however, consideration must be given to the relationship between the test information function and the standard error of estimation. Figure 1 presents this relationship.

Figure 1
Functional Relationship Between Test Information Function and Standard Error of Estimation



As can be seen in this figure, the latter is a strictly decreasing function of the former; yet the amount of decrement in the standard error of estimation is conspicuous for the initial increase of the test information function. It is more or less stabilized, however, after the test information function reaches 20. For instance, for $I(\theta) = 6.25$ the standard error of estimation is .4; this becomes 0.2 (i.e., one-half) when $I(\theta) = 25$. On the other hand, to make the standard error of estimation one-fourth of .4 (i.e., .1), the test information must be 100. This suggests that in tailored testing, increasing the criterion value does not necessarily decrease the standard error of estimation substantially; however, a substantially larger number of items is needed to present to each individual testee. It is desirable, therefore, to find a reasonable balance between the two opposing factors to determine the value of the criterion.

Monte Carlo Study

Method

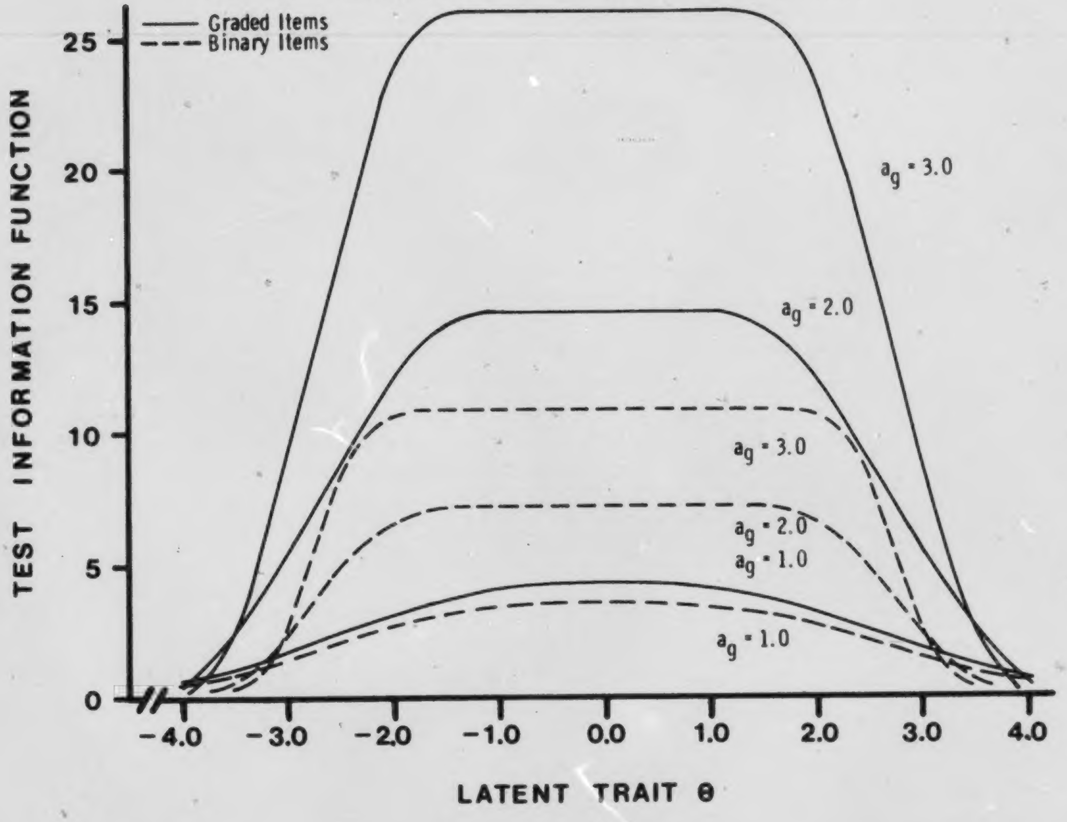
With the considerations previously described, a hypothetical tailored testing situation was constructed using 6 different item pools. The first item pool consisted of 11 types of graded items, each of which had 4 graded item score categories. The difficulty parameters from the normal ogive model for these 11 types of items are presented in Table 1. The other parameter, discrimination index, was 1.0. The second item pool had also 11 types of graded items with the same number of item score categories and values of the difficulty parameters; but the discrimination parameter, a_g, was 2.0 instead of 1.0. The third item pool was the same as the first and the second, except that a_g = 3.0. The other 3 item pools were identical to the first set of 3 item pools, except that the items were binary items and the difficulty parameters were those shown in the column indicated as $x_g = 2$ in Table 1. It was assumed that in each item pool, there was a substantially large number of items of each type so that there were enough items in the hypothetical experiment. Figure 2 shows the test information function of the 11 items of different types taken from each of the 6 item pools.

Table 1
Difficulty Parameters of 11
Hypothetical Graded Items

Item	$x_g = 1$	$x_g = 2$	$x_g = 3$
1	-3.0	-2.5	-2.0
2	-2.5	-2.0	-1.5
3	-2.0	-1.5	-1.0
4	-1.5	-1.0	-0.5
5	-1.0	-0.5	0.0
6	-0.5	0.0	0.5
7	0.0	0.5	1.0
8	0.5	1.0	1.5
9	1.0	1.5	2.0
10	1.5	2.0	2.5
11	2.0	2.5	3.0

Note. The middle values were used as the difficulty parameters of binary items.

Figure 2
Test Information Functions of Six 11-Item Tests



The criterion test information was set as $I(\theta) = 21.63$, the same constant which was used in the monte carlo studies performed in developing the two methods of estimation of the operating characteristics. This value can also be considered as the reasonable compromise suggested in the preceding section; the standard error of estimation was approximately .215. One hundred hypothetical subjects were used in each tailored testing situation. Their ability levels were -2.475 through 2.475 with an interval of .05, similar to the previous two studies of estimation methods. In each pair of tailored testing situations in which the same discrimination parameter was used, the hypothetical group was the same 100 subjects. In practice, the same seed number was used in the two situations to produce the same sequence of random numbers. The first item presented to every subject was Item 6, which is the item with intermediate difficulty. If the subject's score was 0, then the easiest item, Item 1, was presented repeatedly until a score other than 0 was obtained. If the subject's score on Item 6 was m_6, then the most difficult item, Item 11 was presented repeatedly until a score other than m_{11} was obtained. Then the tentative maximum likelihood estimate was computed, and the computer presented an item for which the inforamtion function was greatest at that value of θ. This process was repeated until the test information function at the tentative maximum likelihood estimate exceeded the criterion.

Table 2
Frequency Distribution of the Number of Items Used in Hypothetical Tailored Testing with the Criterion Test Information of 21.63 and Item Discriminations of 1.0

Number of Items	Binary	Graded
27		15
28		74
29		10
30		
31		1
32		
33		
34		
35	1	
36	48	
37	31	
38	9	
39	4	
40	4	
41	2	
42	1	
Total	100	100
Mean	36.92	27.98

Table 3
Frequency Distribution of the Number of Items Used in Hypothetical Tailored Testing with Criterion Test Information of 21.63 and Item Discriminations of 2.0

Number of Items	Binary	Graded
7		21
8		70
9		9
10	1	
11	26	
12	54	
13	10	
14	3	
15	1	
16		
17		
18	1	
Total	96	100
Mean	11.97	7.88

Results

Table 2 presents the frequency distribution of the number of items needed for the hypothetical tailored testing for individual subjects with the criterion 21.63 in each of the two situations where $a_g = 1.0$. A substantial difference between the two frequency distributions can be seen. The mean number of items was 36.92 for the binary case and 27.98 for the graded case, signifying that only 75.8% of the items were necessary in the graded case as compared to the binary case. Tables 3 and 4 show similar results for the two pairs of results in which $a_g = 2.0$ and $a_g = 3.0$, respectively. The mean numbers of items were 11.97 and 7.88 for the cases where $a_g = 2.0$, and 7.38 and 4.56 where $a_g = 3.0$. The respective percentages were 65.8% and 61.8% for these two pairs.

Table 4
Frequency Distribution of the Number of Items Used in Hypothetical Tailored Testing With Criterion Test Information of 21.63 and Item Discriminations of 3.0

Number of Items	Binary	Graded
3		16
4		15
5	2	66
6	2	3
7	55	
8	36	
9	4	
Total	99	100
Mean	7.38	4.56

It can be seen in Tables 2 and 3 that in these two binary cases some subjects were missing. Table 5 explains why there was a failure to obtain maximum likelihood estimates for these five hypothetical subjects. This table shows the initial set of items presented to each subject and the responses. The subjects' tentative maximum likelihood estimates fell in the range of θ where the item information function of any type of item in the respective item pools was practically nil after these initial items had been presented and answered. Thus, there was no way to "move them back" to the range of θ, where the subjects' true individual parameters belonged. These phenomena are closely related to the attenuation paradox in latent trait theory; they are less likely to occur when the discrimination parameters of the items are relatively low in the binary case, and in any graded case, regardless of the discrimination indices (Samejima, 1969).

A question can be raised regarding the goodness of fit of the normal approximation of the error distribution. Since a constant level of test information was used for the criterion, 100 error scores obtained by

$$\varepsilon = \hat{\theta} - \theta \qquad [4]$$

for each group should be distributed approximately $N(0, .215^2)$ if the normal approximation to the conditional distribution of $\hat{\theta}$, given θ, is adequate.

Table 5
Hypothetical Subjects Whose Maximum Likelihood Estimates (MLE) Failed to Converge (Binary Case)

θ	a_g	MLE	Item	Score
-2.475	2.0	-25.40	6	0
			1	0
			1	0
			1	0
			1	0
			1	1
-2.025	2.0	-10.34	6	0
			1	0
			1	0
			1	1
2.325	2.0	10.34	6	1
			11	1
			11	1
			11	0
2.375	2.0	10.34	6	1
			11	1
			11	1
			11	0
2.425	3.0	5.79	6	1
			11	1
			11	1
			11	0

Table 6 presents the results of the chi-square tests for the goodness of fit, using 8 intervals of θ in each case. Considering that $p = .90$ for $\chi^2 = 2.83$ and $p = .10$ for $\chi^2 = 12.02$, with 7 degrees of freedom, this set of 6 test statistics shows a reasonably good fit. In the second half of Table 6, the results of similar chi-square tests can be seen for the 5 groups of 100 hypothetical subjects and for the total group of 500 subjects in the previous two studies of estimating the operating characteristics with a hypothetical paper-and-pencil test of 35 graded items ($m_g = 2$). The distribution of θ for each of these 5 groups was the same as the distribution used in the present study. With 9 degrees of freedom, 6.036 falls between $p = .50$ and $p = .75$.

Discussion and Conclusion

The results in the preceding section strongly support the effectiveness of the graded response item in preference to the binary item, with respect to its branching effect and efficiency in tailored testing. Thus, it can be concluded that in practical applications, it is desirable to include graded response items in adaptive testing item pools.

Table 6
Chi-Square Test for the Goodness of Fit of the Error Frequencies with $N(0, 0.215^2)$

a_g	Binary	*df*	Graded	*df*
1.0	4.621	7	3.367	7
2.0	5.702	7	9.927	7
3.0	11.636	7	5.254	7
Paper-and-Pencil Test of 35 Graded Items				
Total	6.036	9	(500 Subjects)	
Group 1	5.883	7	(100 Subjects)	
Group 2	9.106	7	(100 Subjects)	
Group 3	10.013	7	(100 Subjects)	
Group 4	18.714	7	(100 Subjects)	
Group 5	9.619	7	(100 Subjects)	

These results also indicate that to utilize graded response items effectively in tailored testing, there should be an attempt to develop items with high discriminations. For instance, even with the binary items, reasonably efficient tailored testing can be conducted if $a_g=2.0$, compared with the case in which the items are graded and $a_g=1.0$. Also, the binary case of $a_g=3.0$ is as efficient as the graded case of $a_g=2.0$.

In the tailored testing situation used in the present study, test information was used in terminating the presentation of items. Some questions should be raised concerning this technique. First, the presentation of items is terminated whenever the value of the current test information function exceeds the criterion; however, this information is computed for the current maximum likelihood estimate of the testee's ability, which is almost certainly different from his true ability level. Thus some discrepancy must be expected between this value and the value of the test information function obtained for the testee's true ability level. Secondly, sometimes the test information exceeds the criterion by a substantial amount, especially when the discrimination parameters of the items assume high values. These two facts might affect the goodness of fit of the normal distribution to the error score distribution. The results obtained in the preceding section, however, contradict these conceivable negative effects; and the goodness of fit proved to be quite adequate. There is no reason to doubt that the two methods of estimating the operating characteristics, which were demonstrated for paper-and-pencil tests, can be used in the same way in tailored testing by using a relatively few number of subjects (i.e., as few as several hundred) if the item pool and the procedure of item presentation are well designed. In fact, it is the advantage of tailored testing that constant test information can be achieved easily by using it as the criterion for terminating item presentation. These methods of estimating the operating characteristic will find their full usefulness, therefore, in computerized adaptive testing.

References

Elderton, W. P., & Johnson, N. L. Systems of frequency curves. New York: Cambridge University Press, 1969.

Johnson, N. L., & Kotz, S. Continuous univariate distributions (Vols. 1 and 2). New York: Houghton Mifflin, 1970.

Samejima, F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph, No. 17, 1969.

Samejima, F. The graded response model of latent trait theory and tailored testing. In C. L. Clark (Ed.), Proceedings of the First Conference on Computerized Adaptive Testing. Washington, DC: U.S. Civil Service Commission, 1975, 5-17.

Samejima, F. A use of the information function in tailored testing. Applied Psychological Measurement, 1977, 1, 233-247. (a)

Samejima, F. Weak parallel tests in latent trait theory with some criticisms on classical test theory. Psychometrika, 1977, 42, 193-198. (b)

Samejima, F. A method of estimating item characteristic functions using the maximum likelihood estimate of ability. Psychometrika, 1977, 42, 163-191. (c)

Samejima, F. Comparison of two methods of estimating operating characteristics using the maximum likelihood estimate of latent trait. Paper presented at Psychometric Society Spring Meeting, Chapel Hill, NC, 1977. (d)

DISCUSSION: SESSION 1

R. DARRELL BOCK
THE UNIVERSITY OF CHICAGO

Vale's paper is an interesting application of the multiple nominal categories model to a problem frequent in testing practice, namely, that the test misses the mark in terms of its difficulty level relative to the distribution of ability in the population. Vale's study is concerned with a test that is too easy. It complements an earlier study reported by Kolakowski (1973) of a test that proved to be too difficult for the population for which it was targeted.

The data that Kolakowski reanalyzed came from a field experiment of the effects of the Emergency School Aid program in rural southern schools. Although an effort had been made to select items appropriate for this population, the effort failed; and the majority of the items were in the 20% to 40% difficulty range. As a result, the test did not have a good distribution of scores for purposes of statistical analysis because so many of the students were answering most of the questions incorrectly.

As Vale has done in his study, Kolakowski attempted to improve the score distributions by recovering information from incorrect responses using the nominal model. A result similar, but complementary, to that of Vale's was obtained. A fairly substantial gain in information, relative to binary item scoring, was obtained among low-scoring subjects by utilizing the partial information in their incorrect responses. Many students who had answered all the items incorrectly or were responding near the chance level obtained scores which discriminated among them to some extent. The lower tails of the score distributions were thereby improved for purposes of the subsequent statistical analysis.

The same information gain could probably be obtained from any commercially prepared multiple-choice test administered to a population for which it is too difficult. The detractors in such tests are usually structured to be plausible and partially correct. They tend to elicit responses from persons of low ability that actually are of some value in discriminating among them when scored by means of a model that expresses the probability of response to such detractors as a function of the ability level of the examinees.

Vale is to be congratulated for showing that this type of model can be applied to free-response data in order to recover information from a test that is too easy for the target population. As he pointed out, the multiple-choice format often makes an item much easier than it is under the free-response format. Unfortunately, the free-response format is difficult to accomodate to the machine methods of item scoring and data processing necessary for practical test scoring based on latent trait models. However, technological breakthroughs are possible.

Optical character recognition (OCR) is becoming increasingly reliable and more widely used. It is conceivable that examinees could write their responses in block letters in squares under each item for reading by OCR. If the examinees have been given some practice on responding in this way earlier in the test, it is reasonable to suppose that a high percentage of such responses would be optically readable. Vale's methods of classifying free responses and applying the multiple nominal categories model would then become a practical mode of testing. The multiple-choice item format has always been criticized, and rightly so, for its tendency to elicit clever strategies of outwitting the item writer rather than assessing what the examinee can recall when left to his/her own devices. The procedures demonstrated in Vale's study show that this objection can be overcome.

One question concerns me, however, about the interpretation of Vale's figures: He used as multiple-choice items exactly the same stems that appeared when the item was presented in the free-response mode. Given what we now know about "advance organizers" and cognitive priming, it seems likely that the multiple-choice items were actually easier because they followed the free-response form after a time interval. Mental functioning could be going on during this interval, either consciously or unconsciously. For example, the testee might be unable to think of the meaning of a word immediately upon the first presentation; after ruminating over it or, better yet, forgetting about it entirely, he/she might find that on the next presentation the word would come to mind without effort.

Perhaps the multiple-choice items in this particular experimental design would not have been quite as easy if they had been presented without this kind of priming. This does not, of course, vitiate the principle that Vale's experiment demonstrates; but experiments of this type might be better designed to exclude priming by writing two slightly different forms for each item and assigning one at random to the multiple-choice format and the other to the free-response format.

The use of the misspelling discrimination routine in this study is very clever and appears to have been very effective. This can also be done, as Cory mentioned, by simply having a dictionary of possible misspellings. The dictionary would trap most of the misspellings, and the rest could be assigned to other categories. Nevertheless, the misspelling discrimination routine might be needed if it were desirable to automate free-response scoring in a more ambitious way, for example, to responses longer than one word.

I am quite interested in the tendency demonstrated in the second experiment (which involved 50 items) for the split-half reliability to be slightly greater with conventional scoring than with maximum likelihood scoring. This also occurred with the 9-item test in the first experiment, but because of the small number of items, perhaps not too much should be made of it. After all, maximum likelihood scoring guarantees only asymptotic efficiency; and in this case, the asymptote is with respect to the number of items. Little can be said, therefore, about the results in the case of 9 items; however, in the case of 50 items, it appears that the number-correct score and the maximum likelihood estimate are converging with respect to their reliabilities.

Actually, as the number of items increases, the two forms of scoring become not very different from a monotonic transformation of one another. This is true because the statistic from which the maximum likelihood score is computed is a weighted sum of item responses. As the number of items becomes large, the actual values of the weights become less and less important. For large numbers of items, the number correct score and the maximum likelihood score become essentially equivalent up to a non-linear monotonic transformation; and, as is well known, the Pearson product-moment correlation is not very sensitive to monotonic data transformations.

When comparing number correct scores with maximum likelihood latent trait estimates, however, one must be wary of the tendency of latent trait models based on the normal or logistic response laws to produce outliers. The estimation procedure becomes numerically unstable as there is an attempt to estimate extreme levels of ability, expecially when these ability levels lie outside the difficulty range of items in the test. An example of this appears in the paper by Bock and Kolakowski (1973) in which maximum likelihood estimates of a spatial ability latent trait were resolved into Gaussian components in order to test a genetic hypothesis. To get clear results,a few outliers had to be trimmed from both ends of the distribution of latent trait estimates. In a later paper, Bock and Thrash (1977) performed this resolution without trimming by employing an empirical Bayes solution in which these outliers were allowed to stand but were weighted properly according to the unreliability of scores in that range. Although that solution worked nicely, it involves statistical procedures too complex for most situations in which it would be desirable to use the latent trait estimates as data.

In particular, it is known that the Pearson product-moment correlation coefficient is sensitive to outliers, and any comparison of split-half Pearson product-moment correlations of trait estimates versus number correct scores might be inappropriate. Number correct scores do not have outliers; if anything, their distributions are too short in the tails. Therefore, to compare reliabilities of the latent trait estimates with other types of scores, Spearman rank correlation, which will be robust to the presence of these outliers, should be used.

Clearly,some thought must be given to what to do about the outlier problem in latent trait estimation. I would prefer, wherever possible, to use Samejima's Bayes modal estimate because it quite effectively restrains these extreme values. In this type of estimator, a penalty function is added to the first derivative of the log likelihood and tends to outweigh it as the estimate becomes far removed from the population mean. Since the penalty function tends to have its effect mostly in the tails of the distribution, we can be fairly liberal about accepting the assumption on which it is based, namely, a normal distribution of ability in the population. Even if this assumption is not strictly correct, it will not have much effect in the middle part of the trait distribution or on conclusions drawn from statistics that depend mostly on the middle of the distribution.

Up to this point I have assumed, as the papers in this symposium assume, that the multiple-choice items and free-response items span the same factor space. In view of factor studies such as those of Thurstone and Guilford, it would seem plausible that free responding calls on factors of fluency or creativity to a greater extent than do multiple-choice items. Tests requiring fluent associative recall of information generally show up in these studies as a factor distinct from the two dominant spatial and verbal factors. Detection of minor factors can be quite difficult, but some headway might be made if confirmatory methods of factor analysis rather than exploratory analyses, could be applied. In general, confirmatory analyses will be more powerful in detecting the presence of factors defined in advance than will be exploratory techniques, through which it is hoped that evidence for the factor will appear after the major factors have been accounted for. In particular, it might be possible to build a test based on a facet design that includes free-response versus multiple-choice response as one of its facets. Then a confirmatory factor analysis model could be constructed along the lines suggested by Jöreskog (1969) and tested by the likelihood-ratio methods for the presence of a factor specifically due to the latter facet.

It should always be borne in mind that the latent trait models which have been thus far implemented all assume a unidimensional trait. This assumption should not be made too freely, nor should failure of fit of the latent-trait model be relied upon as an indication of factorial complexity. The tests of fit used by the various programs for latent trait estimation test only whether the characteristic functions are of logistic or normal form. They do not even make use of the information in the joint responses for pairs of items, let alone that in higher-order associations. It has been my experience that a factor analysis of item tetrachoric correlations, for example, will show evidence of more than one factor when the latent trait models (even the very restrictive Rasch model) show good fit as judged by chi-squares computed for either number of correct responses to each item within the score groups or fractiles of latent ability estimates.

Although the item tetrachoric correlation factor analyses do not give a formal test of the number of factors (maximum likelihood factor analysis cannot be used because the tetrachoric correlation matrices are, in general, not positive-definite), the realities of the factors are apparent from the conformity between the factor loadings after promax rotation and the content of the items. In these analyses, the appearance of minor, but real, factors in addition to the main factor represented by the test is quite convincing. Admittedly, the presence of a small number of minor factors is not so important when the normal ogive or logistic ogive model is used to obtain latent trait estimates because the items representing these minor factors will receive small indices of discriminating power. Thus, the items representing the minority factor will have a very small effect on the estimation of the dominant factor. It is important to understand that this is not true of number right scores or of scores based on the Rasch model, because all items are equally influential in determining these types of scores. Therefore, it can happen that scores computed under the normal ogive or logistic model can show surprisingly low correlation with number right scores or scores computed under the Rasch model.

Turning to Cory's paper, I am encouraged to see the concern with extending adaptive testing to other than purely verbal stimuli. Both the results of factor analytic studies and more recent cognitive studies show that there are important mental processes that involve the manipulation of non-verbal images. The kinds of item presentations that Cory discusses open real possibilities for the study of imageal thinking in various modalities. Visual displays and auditory displays are now on the technical horizon.

In Cory's list of job facets, the second most important facet referred to orally presented commands and the understanding of such commands. This is an area that has had very little systematic exploration, even by the cognitive psychologists, because of the lack of well-controlled "displays" of auditory material. What is needed, of course, is some type of economical mechanism for random access of auditory material stored on magnetic disk.

In psychometrics, study of facet designs of this type is still very much at the empirical level. These studies would benefit greatly from experimental set-ups that permitted measurement of response time as well as the correctness or incorrectness of the response. For example, spatial and semantic aspects of recall could be separated by carefully balancing stimuli for a spatial contrast and a semantic contrast, then measuring differences in the amount of response time required for the subject to make a semantic, as opposed to a spatial distinction. In his dissertation, Thissen (1976) showed how information in the correctness of the item response as well as in the response latency can be combined in trait estimation.

The use of graded item scores in these studies should also be encouraged. In personality testing, it is very common to ask the subject to rate the presence of a bipolar trait on a graded scale. Similarly, in developmental studies, the Piagetian stages of development suggest ordered classifications for responses to cognitive tasks. Lieberman's (1970) dissertation, for example, applied Samejima's graded score model to stages of moral development.

In ability testing, it is sometimes less clear how to construct an ordered set of categories reflecting increasing or decreasing ability. In these applications, the nominal model can be quite useful in indicating the order of item categories that is most effective in discriminating people according to their level of ability. The nominal model produces a characteristic curve for each item response category within each item; and by inspecting these curves, the ordering of the categories generally can be seen. Exceptions are those items that are poorly discriminating and have flat characteristic curves for all categories. If a plausible ordering can be found, however, it is then advantageous to use the categories as a graded scale because the graded model requires fewer parameters to be fitted for each item.

Finally, I have a few remarks about Samejima's interesting proposals for estimating item characteristic curves. I understand very much her desire not to ignore the errors of misclassification that must result when maximum likelihood estimates of ability are used to categorize people. I can also understand the desire to keep the number of people required as small as possible. But I tend to be uneasy about the assumption that the distribution of errors in the maximum likelihood estimates about their true value is normal. Again, the justification for this is the asymptotic properties of maximum likelihood estimation; but in this case, the asymptote is with respect to the number of items, which in practical work can never be really large. Kolakowski and I have a paper in progress in which estimates of ability are obtained from monte carlo data in which the true abilities were known. The item parameters chosen were those actually obtained from a set of 38 vocabulary test items. In the first of these analyses, it seemed evident that there was some skewing of the error distribution--that they were not symmetric about the true ability. We are now in the process of rerunning these simulations by independent methods to examine this question more closely. If such asymmetries in the error distribution exist, they would have a disturbing effect on the ability to recover the latent distribution of true ability under the assumption that the measurement error is normally distributed.

My own inclination in attacking this problem would be to proceed iteratively, beginning, for example, with maximum likelihood estimates of abilities based on the normal ogive model. Then the testees could be categorized on the basis of these estimates, and alternative models of response functions could be fitted to the observed percent correct in each of these groups for each of the items. The procedure would be repeated using the new response functions in place of the normal, and so on through a number of iterations. It is hoped that this process would lead to stable estimates of the new class of item characteristic functions. This approach could be expected to be robust in the presence of some error of misclassification of the testees; it would also allow us to work with a side class of response functions.

In this connection, much can be said for a class of response functions based on mixtures of simple response functions. In particular, functions which are weighted sums of simple logistic functions, each with its own intercept and slope parameter, provide a very flexible and mathematically tractable class of curves. Using three-component curves of this type, a wide variety of skewed and short- and long-tailed characteristic functions could be produced. Thissen and I have found that functions of this type are remarkably effective as models for growth curves (Bock & Thissen, 1976; Bock, in press). Inasmuch as there is a close formal similarity between item characteristic curves and growth curves, these models can be exploited when exploring classes of item characteristic curves and methods for fitting such curves. This is an approach that might bear investigation if we find the classes of items for which the conventional normal or logistic item characteristic curves are inappropriate.

References

Bock, R. D. Familial resemblance in patterns of growth in stature. Twin Studies. Proceedings of the Second International Congress on Twin Studies. New York: Alan R. Liss, Inc., in press.

Bock, R. D., & Kolakowski, D. Further evidence of sex-linked major-gene influence on human spatial visualizing ability . The American Journal of Human Genetics, 1973, 25, 1-14.

Bock, R. D., & Thissen, D. Fitting multi-component models for growth in stature. Proceedings of the 9th International Biometric Conference, 1976, 1, 431-442.

Bock, R. D., & Thrash, W. Characterizing a latent trait distribution. In P. R. Krishnaiah (Ed.), Applications of Statistics. Amsterdam: North-Holland Publishing Company, 1977.

Kolakowski, D. Recovering of information in wrong responses. Paper presented at the annual meeting of the American Psychological Association, Montreal, 1973.

Jöreskog, K. G. A general approach to confirmatory maximum likelihood factor analysis. Psychometrika, 1969, 34, 183-202.

Thissen, D. Incorporating item response latencies in latent trait estimation. Unpublished doctoral dissertation, The University of Chicago, 1976.

SESSION 2

ALTERNATIVE MODELS FOR ADAPTIVE TESTING

AN EMPIRICAL EVALUATION OF IMPLIED ORDERS AS A BASIS FOR ADAPTIVE TESTING

NORMAN CLIFF, ROBERT CUDECK, AND DOUGLAS MCCORMICK
UNIVERSITY OF SOUTHERN CALIFORNIA

A MULTIVARIATE MODEL SAMPLING PROCEDURE AND A METHOD OF MULTIDIMENSIONAL TAILORED TESTING

VERN W. URRY
U.S. CIVIL SERVICE COMMISSION

A MODEL FOR TESTING WITH MULTIDIMENSIONAL ITEMS

JAMES B. SYMPSON
UNIVERSITY OF MINNESOTA

DISCUSSION

FREDERIC M. LORD
EDUCATIONAL TESTING SERVICE

SESSION 2: ABSTRACTS

AN EMPIRICAL EVALUATION OF IMPLIED ORDERS AS A BASIS FOR TAILORED TESTING

NORMAN CLIFF, ROBERT CUDECK, AND DOUGLAS MCCORMICK

Implied orders is a simplified model for tailored testing which operates without the requirement of prior knowledge about item parameters; characteristics of items are derived at the same time as persons' responses. The simultaneous mode assumes that a number of examinees are being tested simultaneously; and the sequential, or cumulative, mode that they are tested sequentially with item information used as it accumulates. The simultaneous version was used in monte carlo simulation studies and in a real-data simulation. The sequential version was used in live tailored testing. The results suggest that the sequential version is (1) practical, (2) cost-effective, (3) applicable to small populations, and (4) free of the encumbrances of an item traceline model.

A MULTIVARIATE MODEL SAMPLING PROCEDURE AND A METHOD OF MULTIDIMENSIONAL TAILORED TESTING

VERN W. URRY

A method of multidimensional tailored testing is presented and evaluated in a simulation study; and an asymptotic value of the multiple correlation is sought to provide an optimal level of valid measurement. Terminal standard errors and regression weights are obtained for the several tests that are to be tailored through the use of several item banks. The multidimensional method of tailored testing was used with the data generated by a multivariate model sampling procedure which is designed to maintain representativeness with respect to the joint distributions of abilities and criterion measures through the use of large sample correlation matrices. The obtained results show that very few items were required to attain a value of the multiple correlation comparable to that which would have been obtained through conventional paper-and-pencil testing.

A MODEL FOR TESTING WITH MULTIDIMENSIONAL ITEMS

JAMES B. SYMPSON

Most latent trait models assume that a single latent dimension accounts for the covariation among test items; however, for certain types of items the assumption of unidimensionality is untenable. Existing multidimensional latent trait models are fully compensatory in nature, that is, low status on one latent dimension can be completely offset by high status on other latent dimensions. This characteristic is not appropriate for many test items. A new multidimensional latent trait model that allows only limited compensatory effects is presented. The implications of the partially compensatory model are contrasted with those of the fully compensatory models. An alternating least squares procedure for estimating the parameters of the new model is outlined. Results of a computer simulation of unidimensional parameter estimation using the alternating least squares procedure are presented.

AN EMPIRICAL EVALUATION OF IMPLIED ORDERS AS A BASIS FOR TAILORED TESTING

NORMAN CLIFF
ROBERT CUDECK
DOUGLAS MCCORMICK
UNIVERSITY OF SOUTHERN CALIFORNIA

The tailored testing system described in this paper differs in several respects from others proposed for the same purpose. It does not require thousands of subjects for pretesting as do the tailored testing approaches used by the Educational Testing Service or the Civil Service Commission. It does not require concern about the wisdom of determining item statistics on one population and applying them in another. And it is an appropriate method for integrating testing into a training system.

The basic principle is simple. It arises from considering dichotomous items as furnishing ordering relations between persons and items. If the relations are consistent with each other, considered as a whole they furnish a joint ordering of the persons and the items. It is well known that the logical properties of an order are such that if certain of the relations among elements are known, the remainder can be deduced by making use of the transitivity property which characterizes orders. The basis of this approach to computer-interactive testing has been described by Cliff (1975). The general idea is that even an incomplete matrix of responses by persons to items can be used to deduce some order relations between items, which is their relative difficulty. These order relations in turn can be used to predict what the individual's responses will be to items not yet answered; therefore, the necessity of asking those items would be removed.

Taking a joint order as a model for test items is equivalent to assuming that the data provide a Guttman scale, but test data are not Guttman scales. However, a joint order is only an *approximate* model for test items. The problem in tailored testing is, then, one of modifying the transitivity principle in order to make it work reasonably well in the presence of error. A rough statistical approach is used here. At any given time, there are a certain number of responses implying that item j is harder than item k and a certain number that imply the reverse. If one kind of response predominates over the other, then it is implied that one item is easier than the other. Similarly, the pattern of responses by an individual to a subset of test items may be such that some of the responses imply that he/she should answer a particular test item, as yet untaken, incorrectly. Correspondingly, other responses may imply that he/she should answer it correctly. If one number predominates over the other, then the implication is made correspondingly.

Procedure

Illustrative Examples

Table 1 provides an illustration of the way in which the procedure operates. The two columns on the left show the responses of 15 persons to two items, j and k. To determine which of two items is easier, n_{jk}, the number who answered j correctly and k incorrectly, is examined in comparison to n_{kj}, the number who answered j incorrectly and k correctly. In the data illustrated, Person 5 was the only one who answered j correctly and k incorrectly, whereas Persons 4, 6, 7, 8, 9, 10, and 11 answered in reverse. The Frequencies 1 and 7 (n_{kj} and n_{jk}, respectively) are shown at the bottom of the table. Use of a statistical decision rule, which is outlined below, would lead to the decision that item k was easier than item j. For each pair of items in a test, such a comparison is made by means of the decision rule. The results of the comparisons are recorded in what is called the *item dominance* matrix. In the matrix a "1" means that the row item is more difficult than the column item.

Table 1
Illustrative Basis of TAILOR Process

	Complete				Incomplete		
	Items		Dominant		Items		Dominant
Persons	j	k	Item	Persons	j	k	Item
1	1	1	--	1	1	1	--
2	1	1	--	2	1		--
3	1	1	--	3		1	--
4	0	1	j	4	0		--
5	1	0	k	5		0	--
6	0	1	j	6			--
7	0	1	j	7	0	1	j
8	0	1	j	8	0	1	j
9	0	1	j	9	0		--
10	0	1	j	10			--
11	0	1	j	11	0	1	j
12	0	0	--	12			--
13	0	0	--	13		0	--
14	0	0	--	14	0		--
15	0	0	--	15	0	0	--
	Dominance Frequencies $j = 7$ $k = 1$ $p = 9/256$, therefore $j>k$				Dominance Frequencies $j = 3$ $k = 0$ $p = 1/8$, therefore $j>k$		

The foregoing is applicable to a complete test. In an incomplete or tailored test some of the responses would be missing, as is shown in the two righthand columns of the table. The quantities n_{jk} and n_{kj} can still be counted, however, since Person 5 has now only one item, $n_{kj}=0$. Of the seven persons who answered j incorrectly and k correctly, there is now data on both items from only Persons 7, 8, and 11; so n_{jk} is three. The comparison of n_{jk} to n_{kj} could still lead to the conclusion that item j was more difficult than item k if this were all the information available, provided that the liberal rule were used.

Now consider Person 2, who has answered the more difficult item correctly and has not taken the easier one. It could be concluded that he/she would answer the easier item correctly also, and therefore it would not be administered. Similarly, Person 13 has answered the easier item incorrectly; it could thus be concluded that he/she would answer the more difficult one incorrectly also if he/she were to take it, and it would not be administered. Actually, in making decisions of this kind, what is done is similar to deciding which items are easier and which more difficult. Suppose person i has not yet taken item j. The number of more difficult items he/she has answered correctly would be compared to the number of easier items he/she has answered incorrectly. If, by the same decision rule used earlier, the latter of these were to preponderate over the other, he/she correspondingly would be implied to have also incorrectly answered item i. If the reverse were true, then he/she would be assumed to have answered it correctly. If neither were to preponderate, then no decision would be made. At any given time, then, in the tailored testing process, as many inferences as possible are made about the relative difficulty of the items. These in turn are used to imply responses for each person to items he/she has not yet taken.

Frequency Comparison

A rather liberal two-part decision rule is used in comparing frequencies. The major part corresponds to comparing the frequencies by McNemar's (1969) binomial probability and rejecting the null hypothesis with a one-tailed alpha level of .33. Values of n_{jk} and n_{kj} of 2 to 0 and 3 to 1 thus lead to rejection, and an implication is made. The second aspect of the rule is used to deal with the instances where the frequencies are 1 and 0 only. If the information is very sparse (i.e., early in the testing process), even this small preponderance is used to imply item dominance or an implied response. (This is done by means of a complex probability evaluation which will not be detailed here.) The decisions are thus made on very small frequencies. Although until the end of the testing process a possibility exists that any one of them can be reversed, for the most part they remain quite stable.

In a sense this is not unique; any tailored testing system could fit the above description. What is unusual is the very small frequencies (as few as one) used to make the decisions and the simple decision rule employed. Perhaps the most unusual feature is that the process starts with no knowledge about the items; information on item difficulty is gained at the same rate as knowledge of the individual's abilities is obtained.

Modes of Operation

Group testing. There are two basic modes of operation, which might be termed simultaneous and cumulative. The simultaneous mode, called TAILOR (Cudeck, Cliff, Reynolds, & McCormick, 1976; Cudeck, Cliff, & Kehoe, in press), was developed first. It assumes that a number of subjects are being tested simultaneously with a particular pool of items and that there is no knowledge concerning the items. In the initial round of item presentations, items and persons are randomly assigned to each other for the first pairing. In subsequent rounds, each person is assigned the item that is currently closest to him/her in the joint partial order. The process ends when there is either an actual or an implied response for each person to each item.

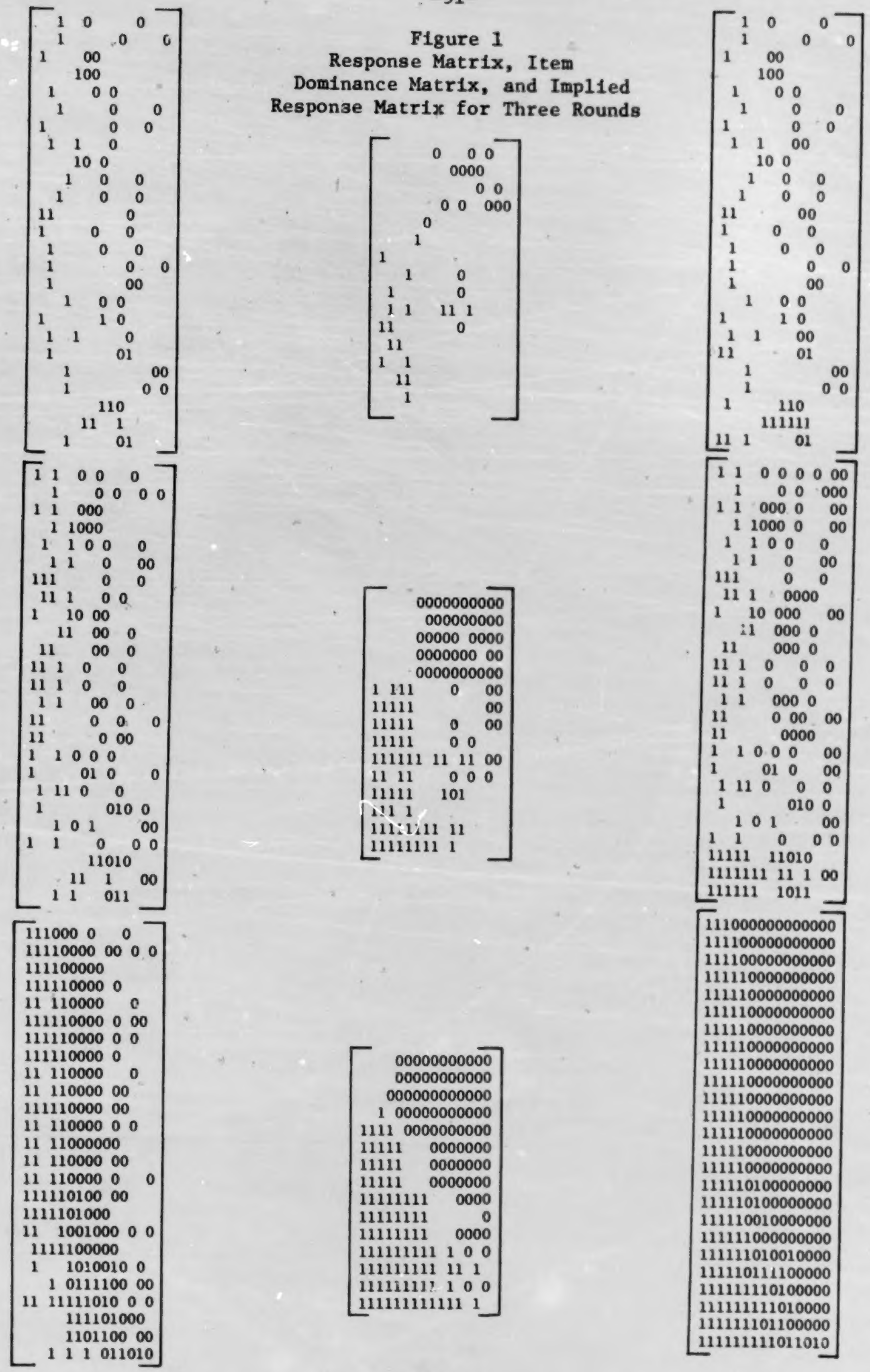
The means of deciding which item to assign to each person is the second major procedure of the process. For its complete operation, this approach carries out one further frequency-comparing step. Each person's implied response vector is compared to every other person's in order to compute a person-person dominance matrix by a means parallel to that used to obtain the item dominance matrix. That is, if person i is implied to answer more of the same items correctly which person h answers incorrectly than items which person h answers correctly and person i answers incorrectly, then i dominates (ranks above) h.

It is possible to assign a current total score to each item and person. For a person, this is the total number of items answered correctly (directly or by implication) minus the number answered incorrectly in the same way, plus the difference in the number of persons he/she dominates and is dominated by. For an item, this score is the number of persons who answer it incorrectly (directly or indirectly) minus the number who answer it correctly, plus the difference in the number of items it dominates and the number which dominate it. In this way items and persons are placed on the same ordinal scale, and the person takes the item for which he/she has no implied or direct response and which is closest to him/her on the scale. Given the various binary matrices involved, this process is actually very simple.

This mode of operation takes place by what might be called "rounds." At each "round," each person should be presented with an item. The item given at one round depends on the results of the previous rounds for that item and person, and each person participates in as many rounds as are necessary to complete his/her score vector.

This procedure is illustrated in Figure 1; the upper three matrices show the operation at any early stage of the process. The data are for 25 persons and 15 items on the Stanford-Binet. The matrices in the left column are the actual response matrices; the middle ones are the item dominance relations that are implied by them; and the right-hand matrices are the implied response matrices. In each a "1" means correct or dominance, a "0" means incorrect or antidominance, and a blank means no relation. The middle set of matrices

Figure 1
Response Matrix, Item Dominance Matrix, and Implied Response Matrix for Three Rounds



shows the operation at an intermediate stage of the testing process, and the bottom set shows the final stage; the matrix in the lower right-hand corner shows that the score matrix is now complete by implication.

Individual testing. The second mode of operation is sequential or cumulative and tests individual subjects only. This is called TAILOR-APL (McCormick & Cliff, in press). Again, no knowledge about the items is assumed. The first person must, therefore, take all items. After a few persons have taken the items, however, there may be enough information to define the relative difficulty of some items. This information is then used to infer the responses of subsequent persons to these items, thereby removing the necessity of taking them. As more and more information accumulates, more relative difficulty relations also accumulate, so that the tests become more and more "tailored" for later subjects.

Data

There are now three kinds of data on one or the other of these methods: (1) monte carlo studies assuming a stochastic model; (2) real-data simulations using data files from complete tests; and (3) actual live tailored testing.

The main dependent variable in each is either the correlation of obtained scores with true scores or correlations between the scores on parallel forms. The major comparisons are with these variables under tailored and complete testing conditions. Of additional interest are a number of variables reflecting cost and efficiency factors and the effects of such elements as statistical parameters of the item pool and circumstances of testing on the results.

Monte Carlo Study

The most extensive data comes from a monte carlo study based on applications of Birnbaum's (1968) three-parameter model. A variety of different characteristics for an item pool were assumed, and a certain number of items were sampled from hypothetical pools with the prescribed characteristics. These item pool characteristics included the mean and standard deviation of the item discrimination parameter, the mean and standard deviation of the item difficulty parameter, and the mean and standard deviation of the item guessing probability. Difficulty and discrimination were assumed to be normally distributed with the prescribed mean and variance; sometimes the variance was zero. It was assumed that a certain number of subjects were being tested simultaneously, and their true scores were sampled from a standard normal population. The program was put in operation and a random number generator was used in conjunction with the Birnbaum model to determine the correctness of each response.

A variety of different combinations of conditions were run; they are given in Table 2. This paper is primarily concerned with a particular subset of conditions. This was a 2 × 2 × 2 factorial design where the variables were (1) number of subjects, 25 or 40; (2) number of items, 15 or 25; and (3) mean discrimination index, a=1.0 or 2.0. Mean item difficulty was zero, as was the variance of item discriminations and the value of the guessing parameter. It should be noted that sampling fluctuations can lead to appreciable

Table 2
Characteristics of Samples of Score Matrices
Generated by Latent Trait Models

		Item Discrimination		Item Difficulty		
Persons	Items	Mean	*S.D.*	Mean	*S.D.*	Mean Guessing
10	25	1	0	0	1	0
10	25	2	0	0	1	0
25	15	.5	0	0	1	0
25	15	.5	0	0	1	0
25	15	1	0	0	1	0
25	15	2	0	0	1	0
25	15	1	0	1	1	0
25	15	2	0	1	1	0
25	15	1	0	0	2	0
25	15	2	0	0	2	0
25	15	1	.2	0	1	0
25	15	2	.2	0	1	0
25	15	2	.4	0	1	0
25	25	1	0	0	1	0
25	25	2	0	0	1	0
25	25	1	0	0	1	.1
25	25	2	0	0	1	.1
25	25	1	0	0	1	.2
25	25	2	0	0	1	.2
25	25	1	.2	0	1	0
25	25	2	.2	0	1	0
25	25	2	.4	0	1	0
40	15	1	0	0	1	0
40	15	2	0	0	1	0
40	15	1	0	0	1	.2
40	15	2	0	0	1	.2
40	25	1	0	0	1	0
40	25	2	0	0	1	0

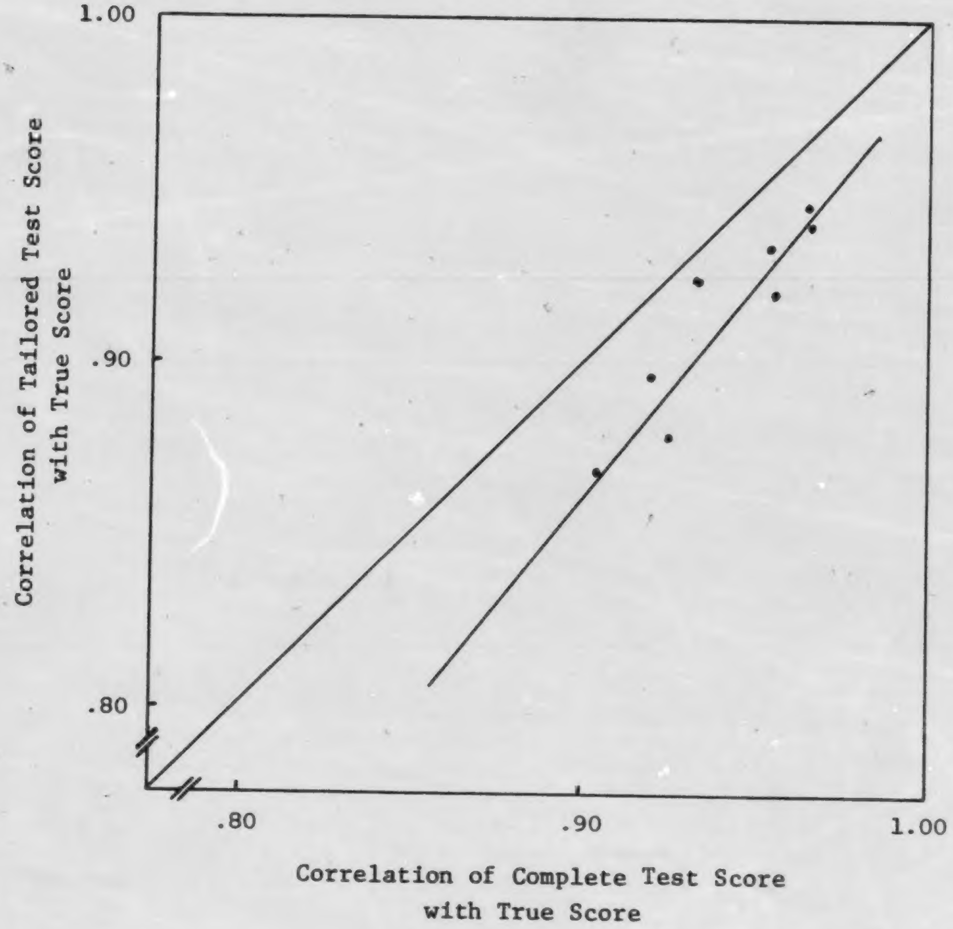
mismatch in difficulty. The average "person" received about half the items under these conditions. More items and/or more persons meant a smaller fraction of items per person, on the average, as shown in Table 3.

Table 3
Proportion of Items Used in Monte Carlo Data

	25 Persons				40 Persons		
	Items				Items		
Item Discrimination	15	25	Mean	Item Discrimination	15	25	Mean
1.0	.609	.514	.562	1.0	.543	.471	.507
2.0	.580	.516	.548	2.0	.548	.441	.494
Mean	.594	.515	.555	Mean	.546	.456	.501

Of more interest, perhaps, is the comparison of the correlation with true score for complete and tailored tests. For these data, the validities were .940 and .913, respectively. Thus, the tailored validities based on half the items were, on the average, 97% as high as those for the complete data. Figure 2 provides more detail, showing a rather close relation between the validity of the complete test and that of the tailored test. Each of the eight points corresponds to the average of five replications of one of the eight combinations of item discrimination parameters, number of items, and number of persons in the 2 × 2 × 2 factorial design. The regression line in Figure 3 may be compared to the 45 degree line which is also indicated. As expected, validity of the tailored test was below that of the complete test; but there was a close correlation. Tailored validity, however, appeared to fall off more steeply than complete validity, i.e., the slope was greater than 1.0.

Figure 2
Relation Between Complete and Tailored Test Validities



Data for all the combinations of conditions showed essentially the same picture: By far the major determinant of tailored validity was complete test validity. Anything that affected the latter (i.e., mean discrimination, guessing probability, number of items in the pool) also affected tailored test validity. Furthermore, the effect was clearly somewhat disproportionate; reducing complete test validity reduced tailored validity even more.

Item File Data

A file of the responses of 625 children from ages 2 to 15 to the Stanford-Binet items was used to simulate the process of actual testing using the simultaneous procedure. The children were divided into three subgroups on the basis of chronological age. Within each subgroup, two samples of 25 items were drawn. A total score was computed on each set of items. One set was used in the TAILOR procedure. The major outcomes of interest were the number of items used in the latter and the correlations of the resulting score with the total score on the untailored half. The simulations were done with either 20 or 40 persons assumed to be tested simultaneously. Within each age group, five samples of 20 and five of 40 were drawn.

The results were quite similar to those for the monte carlo data. About half the items were presented in each case, 55% when 20 persons were tested and 46% when 40 persons were tested. The correlation of the tailored test scores with the complete half-test scores averaged .85 (see Table 4), whereas the correlation between scores on the two complete halves was .88. Thus, the ratio of complete to incomplete correlations was coincidentally again .97. Neither among the age-groups nor between group-sizes were there significant variations. That the latter correlations were somewhat higher in Table 4 is apparently a sampling accident. Of further interest is that responses to 96% of the items not taken were correctly predicted by the procedure.

Table 4
Average Correlations of Tailored
and Complete Tests with a Complete
Parallel Form in Binet Data

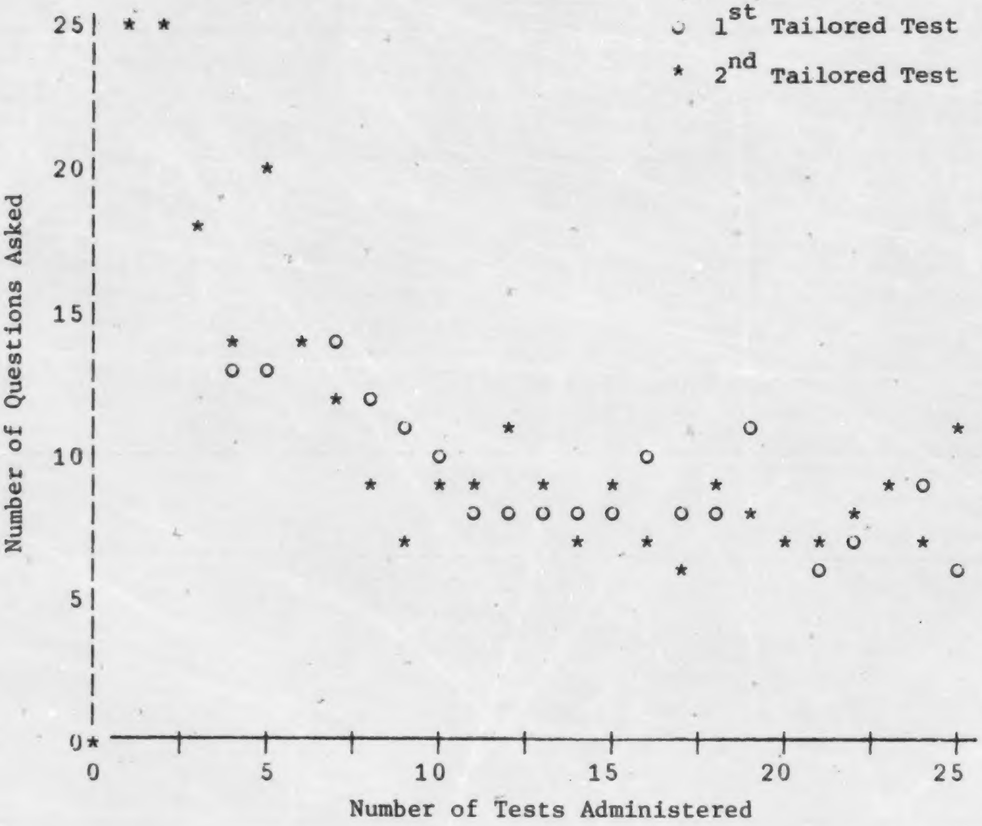
Persons	Comp-Comp	Comp-Tailor
20	.855	.829
40	.889	.866

Live Testing Data

To date, the data with human examinees is only available from the cumulative testing mode, TAILOR-APL. With that procedure only 25 subjects have been used in the tailored and complete testing conditions. The test used was composed of anagrams (scrambled words); it was easy to write a scoring routine for it in a non-multiple-choice mode. It was felt that avoiding a multiple-choice format was desirable on the basis of the results of the monte carlo studies. The computer typed a scrambled word, and the subject's task was to type back the correctly unscrambled word within a specified time limit.

In this mode, too, the items were divided into halves of 25 items each. This time, however, a subject either took two tailored tests simultaneously or two complete tests, rather than one of each as simulated with the Binet data. This procedure may have been important.

Figure 3
Relationship Between Number of Questions Asked and
Number of Tests Administered for First and Second Tailored Tests



It should be remembered that in the cumulative mode, the first subject takes all the items, and each subsequent subject takes progressively fewer items. In Figure 3 the data for the tailored subjects seems to be headed toward an asymptote of around 8 items, even though the average is 11. This given an idea of the rapidity with which this can take place. A different view of the process is presented in Figure 4. Here the box means "correct," the star means "incorrect," and blank means "not taken."

Figure 4
Observed Responses in Each Tailored
Test Arranged Chronologically

```
*******□□□□*□□□□□□□□□□□□□ ******□□**□□□□□□□□□□□□□□□
********□**□*□*□□□□□□□□□□ *********□□***□□*□□□□□□□□
       **********□***□□□□       ** ***********□□□□□
*□□□□□□* □□□□ □           ***□*□□□□□ □□□  □
*****□□* □□□□ □           ******□*□□□□□□□□□□□□
        *  *** *□*□□□*□□□       ******□□□*□□□□
        ******□□□□□□□□          *  ****□□□□□□□
        *****□□□* □□□               * ****□□**
     ** ****□□□□□               *□**□□□
        *□***□□□□ □                 *□*□*□□ □□
     ** ***□□□               * **□□□□*□
     **□**□□□             ****□□□□□ □ □
        *□**□□*□                    *□*□*□□ □□
     ******□□                   **□□□*□
     **□□**□□                       *□**□□□ □□
        ** *□*□ □ □□□           ***□□□□
       ***□*□□□                     ***□□□
     **□*□□□□                *□□□□□ □□*
 *****□□□□ □□                *□□□□* □ □
        ****□□□                     ***□□* □
       **□□*□                   **□□□□□
       *□**□□□                **□□□ *□□
           ****□□□□  □              * ***□**□□
           ****□□*□  □             ***□□□□
       □*□□□*             *****□□□□ □ □
```

The subjects are arranged from earliest to latest, and the two panels show the results for the two subtests. The items in each are ordered from left to right in terms of the final order of difficulty.

Even though an average of half the items were presented, the reliability of the tailored test was substantially, although not significantly, higher than for the complete test (.81 vs. .66 with $N_1=N_2=25$). This is shown quite strikingly in the response matrices for the two conditions, as depicted in Figure 5, where the items are ordered in terms of the difficulty and the persons in terms of scores. The left panel is the incomplete data, the middle panel is the complete matrix inferred from it for the same persons, and the right panel is the corresponding score matrix for the complete data. The two 25-item pools are shown in the upper and lower halves. What appears striking is the substantially greater regularity apparent in the tailored matrix as compared to the complete one. It is much rarer for a person to answer an easy item incorrectly and a difficult one correctly, or vice versa. It appears that the subjects were behaving more consistently under the tailored condition, even though a statistical demonstration of this cannot yet be made.

Figure 5
Response Matrices

```
        TAILORED                  TAILORED                  COMPLETE

        OBSERVED            OBSERVED AND IMPLIED            OBSERVED

*UUUUUU* UUUU U           *UUUUUU*UUUUUUUUUUUUUUUUU **UU**U**UU*UUUUUUUUUUUUU
*****UU* UUUU U           *****UU*UUUUUUUUUUUUUUUUU *U****UUU**UUUUUUU*UUUUUU
 *****UUUU UU             ******UUUUUUUUUUUUUUUUUUU *****UU*UU**UUUU*UUUUUUUU
*******UUUU*UUUUUUUUUUUUU *******UUUU*UUUUUUUUUUUUU ***U***U**UUUU*UUUUUUUUUU
     **U*UUUU             *******U*UUUUUUUUUUUUUUUU ***U*****UUU**UUUUUUUUUUU
     **U**UUU             *******U**UUUUUUUUUUUUUUU ****U*****UUU*UUUUUUUUUUU
     **UU**UU             *******UU**UUUUUUUUUUUUUU ********UU*U*UUUUUUUUUUUU
       U*UUU*             *******U*UUU*UUUUUUUUUUUU ***U****U*UU*UUU*UUUUUUUU
       *U**UUU            ********U**UUUUUUUUUUUUUU ****U*UU*U**U*UUUU*UUUUUU
       **UU*U             *********UU*UUUUUUUUUUUUU **U******U**UUUUUUUUUUUUU
     ** ***UUU            ******* ***UUUUUUUUUUUUUU **U**U******UU*UUUUUUUUUU
     ******UU             ***********UUUUUUUUUUUUUU *****U*U***UU*UU*UUUUUUUU
       ***U*UUU           **********U*UUUUUUUUUUUUU *****UUU****U*U*UUUUUUUUU
********U**U*U*UUUUUUUUUU ********U**U*U*UUUUUUUUUU ****U**UU*****UUUUUUUUUUU
     ** ****UUUUU         ************UUUUUUUUUUUUU *U*********U*U*UUUUUUUUUU
        *U***UUUU U       *********U***UUUUUUUUUUUU *****U***U*U*U**UUUUUUUUU
        *U**UU*U          *********U**UU*UUUUUUUUUU ****U*****UUU*U*UUUU*UUUU
        ****UUU           ************UUUUUUUUUUUUU *****U******U*UUUUUUUUU*U
        ** *U*U U UUU     ********** *U*UUUUUUUUUUU ***********U*UU**UUUUUUUU
        ******UUUUUUUU    **************UUUUUUUUUUU ***********U*U*UUUUU*UUUU
        *****UUU* UUU     *************UUU*UUUUUUUU ***********U*UUUUUU*UU*UU
           ****UUUU  U    ***************UUUUUUUUUU **********U****UU*U*UUUUU
           ****UU*U  U    ***************UU*UUUUUUU **************U*UUUUU*UUU
        *  *** *U*UUU*UUU ****************U*UUU*UUU ******U*U***UU******UUUUU
       **********U***UUUU *****************U***UUUU **************U*U*UUUUUUU

***U*UUUUU UUU  U         ***U*UUUUUUUUUUUUUUUUUUUU ***UU*UUUUUUUUUUU*UUUUUUU
****UUUUU U U             ****UUUUUUUUUUUUUUUUUUUUU *U****UUUU*UU*UUUUUUUUUUU
   *UUUUU UU*             ****UUUUUUUU*UUUUUUUUUUUU *****U*U*UUUUUUUUUUUUUUUU
   *UUUU* U U             ****UUUU*UUUUUUUUUUUUUUUU ***U**U**UU*UUUUUUUUUUUUU
*****UUUU U U             *****UUUUUUUUUUUUUUUUUUUU **U*****UUUUU*UUUUUUUUUUU
******U*UUUUUUUUUUUU      ******U*UUUUUUUUUUUUUUUUU **U****UUU**UUUUUUUUUUUUU
    **UUU *UU             ******UUUU*UUUUUUUUUUUUUU *******UUU*UUUUUUUUUUUUUU
******UU**UUUUUUUUUUUUUUU ******UU**UUUUUUUUUUUUUUU *U*U*U*UU**UU*U*UUUUUUUUU
   * **UUUU*U             *******UUUU*UUUUUUUUUUUUU ******U*UU*UUUUUUUUUUUUUU
      **UUUUU             ********UUUUUUUUUUUUUUUUU ****U****UUUUUUUUUUUUUUUU
      *U**UUU             *******U**UUUUUUUUUUUUUUU *******UUUUU*U*UUU*UUUUUU
      **UUU*U             ********UUU*UUUUUUUUUUUUU **U****UUUUUU**U*UUUUU*UU
      ***UUUU             *********UUUUUUUUUUUUUUUU U***U*U*U**U*U*U*UUUUUUUU
         ***UUUU          ************UUUUUUUUUUUUU ****UUU***U*U*UUU*UUUUUUU
*********UU***UU*UUUUUUUU *********UU***UU*UUUUUUUU ***U*U****U*U*UUUUUUUUUUU
      ******UUU*UUUU      ************UUU*UUUUUUUUU *******UUU*U*U*U*UUUUUUUU
      *  ****UUUUUUU      *************UUUUUUUUUUUU ***U**UU*UU*****UUU**UUUU
          *U*U*UU UU      ***********U*U*UUUUUUUUUU *****U****U**UU*UU*UUUUUU
          *U*U*UU UU      ***********U*U*UUUUUUUUUU U*********U**UU*U**UUUUUU
          *U**UUU UU      ***********U**UUUUUUUUUUU **U****U***U****UU**UUUUU
          ***UUU          *************UUUUUUUUUUUU *****U*******U*U**UUUUUUU
          ***UU* U        *************UU*UUUUUUUUU *******UU*****U**U*UU*UUU
          * ***U**UU      ***************U**UUUUUUU **********U*UUU***U**UUUU
          * ****UU**      ****************UU**UUUUU ********U**U***U**UUU**UU
      ** ***********UUUUU ********************UUUUU ****U*****U*UU**U*U***U*U
```

Discussion

If the appropriate variation on the Spearman-Brown formula is used to compare tailored and complete reliabilities or validities for the simulations, and the formula is solved for the length factor, it appears that the tailored test behaves like a complete test in the simulations with about 25% more responses. This modest saving may be the best that can be done in any system *if pretesting to determine item parameters is included.* Indeed, even from an information-function point of view, it would seem reasonable to conclude that the approach to tailored testing presented here may seem the most plausible. It makes use of the information-function for both items and persons simultaneously, albeit in an informal manner. That is, administering a difficult item to a low ability person does not give much information about either the item *or* the person. Therefore, from the beginning there is an attempt to match items and persons appropriately and thereby obtain more information about both.

Designs Used in Evaluating Tailored Testing

Some type of cross-validation approach is necessary for a tailoring method that relies on prior estimates of item parameters if a reasonable estimate of its efficacy is to be derived. This must be done in a realistic way; spurious correlations between non-independent scores must be avoided. With real data it is at least necessary to estimate parameters on one sample and use them to tailor the test for a second sample. These could be samples from the same population, but it is more realistic to do such things as estimate parameters on data from one year and use it to tailor in a second year. In these ways, the effects of sampling fluctuations in the estimates or item parameters can be more realistically mirrored, since this is how the system would operate in practice.

In the second sample, scores on parallel forms must be available. There must be either two separately tailored and administered tests or a tailored test and a separate conventional test. The parallel form correlation between them must be calculated and then compared to two conventional parallel forms. In this way spurious estimates of agreement are avoided.

If the study is of a monte carlo nature, two samples are still necessary: one for estimating item statistics for use in tailoring and the other to apply them in a tailored test. At present, parallel forms do not seem necessary, since correlations can be presumably calculated with true score and compared with the corresponding correlation for a conventional test. This will give a basis for comparison.

The necessity of pretesting should also be taken into consideration in evaluating efficiency. If 1,000 people take 100 items each in order that a second 1,000 can be given a tailored test of only 20 items, then it seems that the savings due to tailoring are only 40% rather than 80%. Tailoring is only effective if the sample on which item statistics are determined need be only a fraction of that to which the tailored form will be administered. These three considerations--independent samples, independent measures, and inclusion of pretesting costs--seem necessary to be included in the assessment of the usual tailored testing. Of the three, only the necessity for the availability of independent scores seems necessary for the assessment of the approach presented here, however, since the single administration strategy obviates the need for a separate norming sample.

The third point is that the computer may bring psychology back into testing. The subjects seemed to *behave* differently under the tailored condition, perhaps their minds even worked differently. If the data are taken at face value, the eleven tailored responses acted like a 55-item complete test. It appears that the subjects simply behaved more consistently on the tailored test; a high ability subject was less likely to give an incorrect answer to an easy item and a low ability subject was less likely to correctly answer a difficult item.

Cost

The monte carlo studies cost about an average of ten cents per pseudo-subject; the efficiency of the program can probably still be increased by a factor of two to five, and computer costs will reduce by a similar factor in the next three or four years. Therefore, this aspect of the cost of testing will be relatively small in most applications compared with, for example, item writing. The actual computer costs of administering the anagrams test to one person was about $3.50; current revisions of the program should reduce this to below $1.00.

Conclusions

The tailored testing procedure described in this paper appears to be cost-effective, applicable to small populations, and relatively free from the encumbrances of an item traceline model. Thus, it might be a viable alternative to other approaches to tailored or adaptive testing.

References

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), Statistical theories of mental test scores (Part 5). Reading, MA: Addison-Wesley, 1968.

Cliff, N. Complete orders from incomplete data: Interactive ordering and tailored testing. Psychological Bulletin, 1975, 82, 289-302.

Cudeck, R., Cliff, N., & Kehoe, J. TAILOR: A FORTRAN program for interactive tailored testing. Educational and Psychological Measurement, in press.

Cudeck, R., Cliff, N., Reynolds, T., & McCormick, D. Monte carlo results from a computer program for tailored testing (Technical Report No. 2). University of Southern California, Department of Psychology, 1976.

McCormick, D., & Cliff, N. TAILOR-APL: An interactive program for individual tailored testing. Educational and Psychological Measurement, in press.

McNemar, Q. Psychological statistics (4th ed.). New York: Wiley, 1969.

Acknowledgements

This research was supported by the Office of Naval Research, Contract N00014-75-C-0684, NR 150-373. Appreciation is extended to Mark Reckase of the University of Missouri for making available the item file data from the Stanford-Binet. The live testing data are from a Master's Dissertation written by Douglas McCormick.

A MULTIVARIATE MODEL SAMPLING PROCEDURE AND A METHOD OF MULTIDIMENSIONAL TAILORED TESTING

VERN W. URRY
U.S. CIVIL SERVICE COMMISSION

It has been established that in the unidimensional case in which only one ability is tested (Jensema, 1974, 1976; Urry, 1974, 1975, 1977), computer-assisted tailored testing can be very economical. The reliabilities typical of paper-and-pencil tests of single abilities can be achieved with far fewer items. This economy results from the computer-interactive adaptation of the item sequence to the level of ability of the particular examinee. In one study (Urry, 1977) the number of items required to achieve a specified reliability was only one-fifth as great as the number required by a conventional paper-and-pencil test: The paper-and-pencil test required 100 questions and the tailored test required only 20.

When several calibrated item banks are available for the measurement of several abilities, the problem becomes one of determining the most economical use of these banks to optimize the multidimensional validity of a composite of the several tailored test scores. The conventional paper-and-pencil test analogue of this problem is to differentially allocate the number of items to the various tests in a battery so that the multiple correlation is maximized for a fixed testing time. This particular problem was addressed earlier by Taylor (1939, 1950), Horst (1949, 1956), and Woodbury and Novick (1968).

The problem is addressed in the present paper from the perspective of tailored testing. The Woodbury and Novick solution is modified to provide--at an asymptotic value of the maximum multiple correlation as a function of testing time--allocations by item banks for (1) terminal reliabilities, (2) terminal standard errors, and (3) appropriate weights for ability estimates obtained from tailored testing. The modified solution can be used in conjunction with the Owen unidimensional algorithm. The result is a multidimensional algorithm appropriate for use when (1) tests are tailored from several item banks for each examinee and (2) an external measure of job proficiency is available.

In the Owen algorithm, tailored testing can be terminated when a specific value of the standard error of the ability estimate is achieved. The specific value for a given bank is referred to as a terminal standard error. In actuality, this is the square root of the Bayesian posterior variance. Since the standard deviation of ability has been set equal to unity, 1.0 minus the Bayesian posterior variance will yield the terminal reliability when tailored testing has been terminated at a specific terminal standard error. A composite score, or predicted criterion score, can then be obtained by merely adding tailored test scores (ability estimates) that have been multiplied by their appropriate weights. The appropriate weights in this context are regressed score weights. The extent of regression, or the standard deviation of the regressed scores, is given by the square root of the terminal reliability. This is the construct validity coefficient or the slope of the regression of tailored test scores (ability estimates) on true ability. True ability, through calibration, has a standard deviation of 1.0.

In the balance of this paper, the modified Woodbury and Novick solution is detailed, a multivariate item response generator is described which generated data for a simulation study, the design of the simulation study to assess the effectiveness of the multidimensional algorithm is presented, results obtained from the simulation study are given, and the important implications of the multidimensional algorithm for tailored testing are considered.

Method

Multivariate Item Response Generation

In multivariate item response generation, several true ability scores and a true criterion score are sampled for each simulated examinee. Given these true ability scores, binary item responses (that is, zeroes or ones indicating incorrect or correct answers) are sampled for the items. For convenience, the item responses for each simulated examinee are arranged on the basis of the particular ability each item measures. These data are then available for the simulation of multidimensional tailored testing. Each ability can be estimated through a unidimensional tailoring algorithm using the appropriate item responses; each ability estimate can then be compared with its corresponding true value. In addition, appropriate weights can be applied to each ability estimate to obtain a composite or predicted criterion score that can, in turn, be compared with a true criterion parameter.

An estimate of the population supermatrix P is required. This symmetric supermatrix has the following partitioned structure:

$$P = \left[\begin{array}{c|c} P_{\theta\theta} & \underline{\rho} \\ \hline \underline{\rho}' & 1.00 \end{array}\right] = \left[\begin{array}{cccc|c} 1.00 & \rho_{\theta_1\theta_2} & \cdots & \rho_{\theta_1\theta_p} & \rho_{\theta_1 y} \\ \rho_{\theta_2\theta_1} & 1.00 & \cdots & \rho_{\theta_2\theta_p} & \rho_{\theta_2 y} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho_{\theta_p\theta_1} & \rho_{\theta_p\theta_2} & \cdots & 1.00 & \rho_{\theta_p y} \\ \hline \rho_{y\theta_1} & \rho_{y\theta_2} & \cdots & \rho_{y\theta_p} & 1.00 \end{array}\right] \quad [1]$$

where $P_{\theta\theta}$ is the matrix of intercorrelations between the latent abilities, θ_k, for $k = 1, 2 \ldots p$;

$\underline{\rho}$ is a column vector of correlations $\rho_{\theta_k y}$ between the latent abilities θ_k and a criterion variable, y; and

$\underline{\rho}'$ is the transpose of $\underline{\rho}$, or a row vector of the validity coefficients for the latent abilities.

While the supermatrix P is never observed in practice,[1] it can be satisfactorily estimated from supermatrices of attenuated correlations based on large samples through the use of

$$\mathrm{P} = D_r^{-\frac{1}{2}} \left[R-I\right] D_r^{-\frac{1}{2}} + I. \qquad [2]$$

In Equation 2, the matrix I is the $(p+1)$ by $(p+1)$ identity matrix. The supermatrix R is partitioned as follows:

$$R = \left[\begin{array}{c|c} R_{\hat\theta\hat\theta} & \underline{r} \\ \hline \underline{r}' & 1.00 \end{array}\right] = \left[\begin{array}{cccc|c} 1.00 & r_{\hat\theta_1\hat\theta_2} & \cdots & r_{\hat\theta_1\hat\theta_p} & r_{\hat\theta_1\hat y} \\ r_{\hat\theta_2\hat\theta_1} & 1.00 & \cdots & r_{\hat\theta_2\hat\theta_p} & r_{\hat\theta_2\hat y} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ r_{\hat\theta_p\hat\theta_1} & r_{\hat\theta_p\hat\theta_2} & \cdots & 1.00 & r_{\hat\theta_p\hat y} \\ \hline r_{\hat y\hat\theta_1} & r_{\hat y\hat\theta_2} & \cdots & r_{\hat y\hat\theta_p} & 1.00 \end{array}\right] \qquad [3]$$

where $R_{\hat\theta\hat\theta}$ is the matrix of attenuated intercorrelations between less than perfectly reliable or fallible measures of the latent abilities, $\hat\theta_k$, for $k = 1, 2 \ldots p$;

$\underline{r}$ is a column vector of attenuated correlations, $r_{\hat\theta_k\hat y}$, between the fallible measures of latent abilities, $\hat\theta_k$, and a fallible criterion variable, $\hat y$; and

$\underline{r}'$ is the transpose of $\underline{r}$, or a row vector of validity coefficients attenuated in the variables.

---

[1] The supermatrix P represents the intercorrelations between perfectly reliable ability and criterion measures. While this supermatrix exists in theory, in practice, perfectly reliable ability and criterion measures are exceptional.

The supermatrix $D_r^{-\frac{1}{2}}$ is partitioned as follows:

$$D_r^{-\frac{1}{2}} = \left[\begin{array}{c|c} D^{-\frac{1}{2}}_{r_{\hat\theta\hat\theta}} & \underline{0} \\ \hline \underline{0}' & \frac{1}{\sqrt{r_{\hat y\hat y}}} \end{array}\right] = \left[\begin{array}{ccccc|c} \frac{1}{\sqrt{r_{\hat\theta_1\hat\theta_1}}} & 0 & \cdots & 0 & & 0 \\ 0 & \frac{1}{\sqrt{r_{\hat\theta_2\hat\theta_2}}} & \cdots & 0 & & 0 \\ \cdots & \cdots & \cdots & 0 & & \cdots \\ 0 & 0 & \cdots & \frac{1}{\sqrt{r_{\hat\theta_p\hat\theta_p}}} & & 0 \\ \hline 0 & 0 & \cdots & 0 & & \frac{1}{\sqrt{r_{\hat y\hat y}}} \end{array}\right] \qquad [4]$$

where $D^{-\frac{1}{2}}_{r_{\hat\theta\hat\theta}}$ is a p by p diagonal matrix of reciprocal square roots of the reliabilities of the fallible measures of the latent abilities;
$\underline{0}$ is a column or p by 1 null vector;
$\underline{0}'$ is the transpose of $\underline{0}$ or a row or 1 by p null vector; and
$\frac{1}{\sqrt{r_{\hat y\hat y}}}$, a scalar, is the reciprocal square root of the reliability of the criterion variable.

The estimate of the supermatrix P is decomposable into its eigenvectors and eigenvalues. This process yields the identity

$$\hat{P} = QDQ' , \qquad [5]$$

where Q is the $(p+1)$ by $(p+1)$ matrix of the eigenvectors of $\hat{P}$;
D is the $(p+1)$ by $(p+1)$ diagonal matrix of the eigenvalues of $\hat{P}$ in descending order of magnitude; and
Q' is the transpose of Q.

A $(p+1)$ by $(p+1)$ matrix of weights

$$W = D^{\frac{1}{2}}Q' \qquad [6]$$

is obtained for later use, where the diagonal matrix $D^{\frac{1}{2}}$ contains the square roots of the eigenvalues in descending order of magnitude and the matrix Q' is as previously defined. A matrix T can then be obtained through

$$T = ZW, \qquad [7]$$

where Z is the N by $(p+1)$ matrix, the elements of which are merely independent, drawing from the normal distribution, $N(0,1)$, with a mean of zero and a variance of one; and the matrix W is as defined in Equation 6. The matrix T is partitioned as follows:

$$T = \left[\theta \,\middle|\, \underline{y}\right] = \left[\begin{array}{cccc|c} \theta_{11} & \theta_{12} & \cdots & \theta_{1p} & y_1 \\ \theta_{21} & \theta_{22} & \cdots & \theta_{2p} & y_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \theta_{ij} & \cdots & y_j \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \theta_{N1} & \theta_{N2} & \cdots & \theta_{Np} & y_N \end{array}\right] \qquad [8]$$

where θ is the N by p matrix of ability parameters in standard score form, and $\underline{y}$ is the N by 1 column vector of criterion parameters in standard score form.

Since the expectation, E, of $\frac{Z'Z}{N}$ is equal to the identity matrix, i.e.,

$$E\left[\frac{Z'Z}{N}\right] = I \,, \qquad [9]$$

it then follows that the expectation, E, of $\frac{T'T}{N}$, or

$$E\left[\frac{T'T}{N}\right] = QD^{\frac{1}{2}}\, E\left[\frac{Z'Z}{N}\right] D^{\frac{1}{2}}Q' = \hat{P}, \qquad [10]$$

is equal to the desired supermatrix. Thus, the operation in Equation 7 provides a simulated sample of size N from the population defined by $\hat{P}$.

Let it be assumed that representative values for the item parameters are available for items measuring the several abilities. For example, the i^{th} item measuring the k^{th} ability would have the parameters $\underline{a}_{ik}$ (discriminatory power), b_{ik} (difficulty), and c_{ik} (coefficient of guessing), where $i = 1, 2, \ldots q_k$, the number of items measuring the k^{th} ability; q_k is the last item in the k^{th} bank; and $k = 1, 2, \ldots p$, the number of abilities measured.

Given θ_{jk} (examinee j's true ability score on ability k) and the parameters for item i on ability k, the j^{th} simulated examinee's binary response, or u_{ijk}, (that is, 0 or 1 indicating an incorrect or correct answer) is obtained through evaluating

$$P'_{ik}(\theta_{jk}) = c_{ik} + (1-c_{ik})P_{ik}(\theta_{jk}) \,, \qquad [11]$$

where $P'_{ik}(\theta_{jk})$ is the proportion obtaining a binary score of 1 at θ_{jk}, and

$P_{ik}(\theta_{jk})$, the proportion knowing the correct answer (as opposed to knowing or guessing correctly) at θ_{jk} is

$$P_{ik}(\theta_{jk}) = \left[1 + \exp\{-Da_{ik}(\theta_{jk} - b_{ik})\}\right]^{-1}, \qquad [12]$$

where D is the constant 1.7. Because of the complementary relationship,

$$Q_{ik}(\theta_{jk}) = 1 - P'_{ik}(\theta_{jk}) , \qquad [13]$$

the exhaustive and mutually exclusive events (that is, 0 or 1 indicating an incorrect or correct answer to the i^{th} item measuring the k^{th} ability) may be mapped onto the unit interval.

Thereafter, a random number, r_u, is drawn from a distribution of uniform density on the interval from 0 to 1. Given that

$$r_u > P'_{ik}(\theta_{jk}) , \qquad [14]$$

assign

$$u_{ijk} = 0 \text{ (incorrect)} . \qquad [15]$$

Otherwise, or when

$$r_u \leq P'_{ik}(\theta_{jk}) , \qquad [16]$$

assign

$$u_{ijk} = 1 \text{ (correct)} . \qquad [17]$$

The process is merely repeated for distinct u_{ijk}.

For the purpose of subsequent processing, it is convenient to structure a complete record for the j^{th} simulated examinee in the following manner:

$$\begin{array}{lllllll}
\theta_{j1} & u_{1j1} & u_{2j1} & \cdots & u_{ij1} & \cdots & u_{q_1 j1} \\
\theta_{j2} & u_{1j2} & u_{2j2} & \cdots & u_{ij2} & \cdots & u_{q_2 j2} \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\theta_{jk} & u_{1jk} & u_{2jk} & \cdots & u_{ijk} & \cdots & u_{q_k jk} \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\theta_{jp} & u_{1jp} & u_{2jp} & \cdots & u_{ijp} & \cdots & u_{q_p jp} \\
y_j & & & & & &
\end{array}$$

where θ_{jk} is the parameter in standard form on the k^{th} ability for the j^{th} examinee;

u_{ijk} is the binary score for the j^{th} examinee on the i^{th} item measuring the k^{th} ability; and

y_j is the criterion parameter in standard form for the j^{th} examinee.

For convenience, the number of the last item in the k^{th} item bank, q_k, is illustrated in this record as constant across banks. It need not be held constant in practice.

Modified Woodbury and Novick Solution

The developments in this portion of the paper parallel those given by Woodbury and Novick (1968), with some notational change. Where simplifications were possible and modifications necessary, these are introducted.

The investigators designate a diagonal matrix D with main diagonal elements d_{kk} given by

$$d_{kk} = \frac{\sqrt{t_k}}{\sqrt{\sigma^2_{x_k(t_k)}\left[1-\rho_{x_k(t_k)x'_k(t_k)}\right]}} , \qquad [18]$$

where t_k is the time allocated to the k^{th} test in the battery. The variance and reliability of the k^{th} test are $\sigma^2_{x_k(t_k)}$ and $\rho_{x_k(t_k)x'_k(t_k)}$, respectively, for the allocated time, t_k. Thus, the total time allocated to the n tests in the battery is given by

$$T = \sum_{k=1}^{n} t_k . \qquad [19]$$

In further developments, a product of matrices is useful:

$$D\Sigma D = \left[D^* D^{-\frac{1}{2}}_{diag\ \Sigma}\right] \Sigma \left[D^{-\frac{1}{2}}_{diag\ \Sigma}\ D^*\right] = D^* R D^* , \qquad [20]$$

where the elements of D were defined in Equation 18, and Σ is the variance-covariance matrix for the tests under the allocated times that sum to T. In actual practice, it will be convenient to work with the righthand equality, where R (the intercorrelation matrix for the tests under the allocated times that sum to T)is given by

$$R = D^{-\frac{1}{2}}_{diag\ \Sigma}\ \Sigma\ D^{-\frac{1}{2}}_{diag\ \Sigma} , \qquad [21]$$

as indicated by the equalities in Equation 20 in which $D^{-\frac{1}{2}}_{diag\ \Sigma}$ is a diagonal matrix of reciprocal square roots of the diagonal elements (test variances) of $\boldsymbol{\Sigma}$. Since D is the product of D^* and $D^{-\frac{1}{2}}_{diag\ \Sigma}$, D^* may be observed to have diagonal elements

$$d^*_{kk} = \frac{\sqrt{t_k}}{\sqrt{1 - \rho_{x_k(t_k)x'_k(t_k)}}}, \qquad [22]$$

where the terms are as previously defined. The matrix

$$F = D_t^{-1}(D^*RD^* - D_t)\ D_t^{-1} \qquad [23]$$

is then obtained for future use. In this equation, D_t is a diagonal matrix with elements t_k (as previously defined) on its main diagonal. The diagonal matrix D_t^{-1} contains, by convention, the reciprocals of t_k on its main diagonal.

It is also required that there be a column vector

$$D\ \underline{Cov}_{x(t)y} \Big/ \sqrt{\sigma^2_y} = D^*D^{-\frac{1}{2}}_{diag\ \Sigma}\ \underline{r}_{x(t)y}\sigma_y \Big/ \sqrt{\sigma^2_y} = D^*\underline{r}_{x(t)y}, \qquad [24]$$

where D, D^*, $D^{-\frac{1}{2}}_{diag\ \Sigma}$ are as previously defined;

$\underline{Cov}_{x(t)y}$	is a column vector of covariances, $r_{x_k(t_k)y}\ \sigma_{x_k(t_k)}\ \sigma_y$, between the tests, under the allocated times that sum to T, and the criterion y;
σ_y and σ^2_y	are the standard deviation and variance, respectively, for the criterion y; and
$\underline{r}_{x(t)y}$	is a column vector of validity coefficients under the allocated times that sum to T.

Again, in actual practice, it is convenient to work with the righthand equality in Equation 24. For future use, a column vector

$$\underline{\Upsilon} = D_t^{-1}\ D^*\ \underline{r}_{x(t)y} \qquad [25]$$

is obtained where all terms have been previously defined.

A valid solution requires that all the optimally allocated testing times, t^*_k, in the column vector

$$\underline{t}^* = \frac{T^* + \underline{e}'\ F^{-1}\ \underline{e}}{\underline{e}'\ F^{-1}\underline{\Upsilon}}\ F^{-1}\underline{\Upsilon} - F^{-1}\underline{e} \qquad [26]$$

be non-negative. In Equation 26, T^* is the total time available for testing, i.e.,

$$T^* = \sum_{k=1}^{n} t_k^* \qquad [27]$$

where the column vector $\underline{e}$ is an elementary vector of length $\underline{n}$, the elements of which are all unity;

$\underline{e}'$, the transpose of $\underline{e}$, is a row vector of length n, the elements of which are all unity;

F^{-1} is the inverse of the output matrix of Equation 23; and

$\underline{Y}$ is the output vector of Equation 25.

An application is begun with a large value for T^*, which is then decreased systematically. When a negative t_k^* is encountered, the k^{th} test is dropped from the solution by appropriately reducing the involved matrices, vectors, and scalars. In this case, the terminal reliability, the terminal standard error, and the weight for regressed ability estimate on the k^{th} test are set to zero, unity, and zero, respectively; and the original subscripting is preserved for the purpose of tailored testing.

A diagonal matrix D_{t^*} is defined to contain the elements of $\underline{t}^*$ of Equation 26 on its main diagonal. The diagonal matrix then required is

$$\Lambda = \left[I + (D_{t^*} D_t^{-1} - I)\, D_r \right] D_{t^*}^{-1} D_t \qquad [28]$$

where I is the identity matrix;

D_{t^*}, D_t, D_t^{-1} are as previously defined;

D_r is a diagonal matrix containing the reliabilities of the tests under the allocated times that sum to T on its main diagonal; and

$D_{t^*}^{-1}$ is the inverse of D_{t^*}.

At a particular T^*, the terminal reliabilities for tailored testing are then given on the main diagonal of

$$D_{\tilde{r}} = \Lambda^{-1} D_r\,, \qquad [29]$$

where Λ^{-1} is the inverse of the diagonal matrix defined in Equation 28. It is of interest to interpret a diagonal element of Equation 29. From Equations 28 and 29, it can be deduced that a diagonal element of $D_{\tilde{r}}$, namely $\tilde{r}_{kk}$, is defined as

$$\tilde{r}_{kk} = \frac{\dfrac{t_k^*}{t_k}\, r_{kk}}{1 + \left(\dfrac{t_k^*}{t_k} - 1\right) r_{kk}}\,, \qquad [30]$$

which is the continuous form of the Spearman-Brown formula, where $\dfrac{t_k^*}{t_k}$ is the continuous analogue of the discrete integer k as given in

$$\tilde{r}_{kk} = \frac{k\ r_{kk}}{1 + (k - 1)\ r_{kk}} \qquad [31]$$

under the usual notation for the formula. The diagonal elements of $D_{\tilde{r}}$ are thus the appropriately altered reliabilities from a solution for a particular T^*.

At a particular T^*, the terminal standard errors for tailored testing are given on the main diagonal of

$$D_{\tilde{\sigma}_{\varepsilon}} = \left[I - D_{\tilde{r}}\right]^{\frac{1}{2}} \quad , \qquad [32]$$

where I is the identity matrix and $D_{\tilde{r}}$ is defined in Equation 29. Thus, a diagonal element of $D_{\tilde{\sigma}_{\varepsilon}}$, namely $\tilde{\sigma}_{\varepsilon_k}$, can be interpreted from Equation 32 as

$$\tilde{\sigma}_{\varepsilon_k} = \sqrt{1 - \tilde{r}_{kk}} \quad , \qquad [33]$$

the square root of 1 minus the reliability.

The squared maximum multiple correlation for the weighted composites of ability estimates, R_c^2, for a particular T^* is given by

$$R_c^2 = \underline{r}'_{x(t)y}\ (R + \Lambda - I)^{-1}\ \underline{r}_{x(t)y} \quad , \qquad [34]$$

where $\underline{r}'_{x(t)y}$, a row vector, is the transpose of $\underline{r}_{x(t)y}$ as previously defined in Equation 24;
R and Λ are as previously defined in Equations 20 and 28, respectively; and
I is the identity matrix.

It is necessary to derive the appropriate weights for regressed ability estimates at a given T^*. Standard weights, $\tilde{\beta}_k^*$, are obtained through the normal equations provided by calculus,

$$\tilde{\underline{r}}_{x(t^*)y} = \tilde{R}\tilde{\underline{\beta}}^* \quad , \qquad [35]$$

where $\tilde{\underline{r}}_{x(t^*)y}$ and $\tilde{R}$ are the appropriately altered column validity vector and intercorrelation matrix for the tests in which the allocations of time sum to T^*. The altered column validity vector is known from Woodbury and Novick (1968) to be given by

$$\tilde{\underline{r}}_{\tilde{x}(t^*)y} = \Lambda^{-\frac{1}{2}} \underline{r}_{x(t)y} \quad , \qquad [36]$$

and the altered intercorrelation matrix is similarly known to be given by

$$\tilde{R} = \Lambda^{-\frac{1}{2}}\left[R + \Lambda - I\right]\Lambda^{-\frac{1}{2}} \quad . \qquad [37]$$

Thus, an explicit solution for $\tilde{\underline{\beta}}^*$ involves the premultiplication of both sides of Equation 35 by $\tilde{R}^{-1}$. Then

$$\tilde{\underline{\beta}}^* = \tilde{R}^{-1}\tilde{\underline{r}}_{\tilde{x}(t^*)y} \quad . \qquad [38]$$

But it is known that

$$\tilde{R}^{-1} = \Lambda^{\frac{1}{2}}\left[R + \Lambda - I\right]^{-1}\Lambda^{\frac{1}{2}} \quad , \qquad [39]$$

because the inverse of a product of square basic matrices is equal to the product of their inverses in reverse order. Substitution from Equations 36 and 39 into Equation 38 now provides a more convenient form,

$$\tilde{\underline{\beta}}^* = \Lambda^{\frac{1}{2}}\left[R + \Lambda - I\right]^{-1}\underline{r}_{x(t)y} \quad . \qquad [40]$$

The weights for regressed ability estimates appropriate in tailored testing are now obtained from

$$\tilde{\underline{b}} = D_{\tilde{r}}^{-\frac{1}{2}}\underline{\beta}^* \quad , \qquad [41]$$

because the main diagonal elements of $D_{\tilde{r}}^{-\frac{1}{2}}$ are the reciprocal square roots of the reliabilities or the reciprocal standard deviations of the regressed ability estimates provided by the Owen algorithm. A predicted criterion score, $\hat{y}_j$, is obtained with

$$\hat{y}_j = \hat{\underline{\theta}}'\tilde{\underline{b}} = \hat{\underline{\theta}}' D_{\tilde{r}}^{-\frac{1}{2}}\tilde{\underline{\beta}}^* \quad , \qquad [42]$$

where $\hat{\underline{\theta}}'$ is a row vector of regressed ability estimates from the Owen algorithm (one for each ability bank) and the remaining terms are as previously defined. The middle equality in Equation 42 is the most convenient form, but the righthand equality is more informative. In the righthand equality, the product $\hat{\underline{\theta}}' D_{\tilde{r}}^{-\frac{1}{2}}$ can be seen to standardize ("unregress") the regressed ability estimates prior to the application of standard weights; concomitantly, this product can be viewed as an operation that unbiases the regressed ability estimates or renders them on the same scale as the corresponding true abilities.

Of interest, in actual practice, are the asymptotic properties of the maximum multiple correlation as T^* increases. Beyond some point on T^*, increased testing time yields diminishingly small increases in validity, as indexed by the maximum multiple correlation. A solution at a specific T^*

is then selected in which negligible increase in the maximum multiple correlation is expected with an increase in testing time. The terminal reliabilities, terminal standard errors, and appropriate weights for tailored testing may then be obtained for the selected value of T^*.

Design of the Simulation Study

The population intercorrelation matrix, by assumption, contains the intercorrelations of the latent abilities, their correlations with a criterion, and the criterion self-correlation, unity. To assure verisimilitude, this matrix of parameters should be an intercorrelation matrix actually obtained in a large sample and later disattenuated in the tests. A matrix of parameters implies error-free tests. An attenuated matrix was obtained from French (1963). This matrix is well known because it was also used in the six-predictor-variable problem analyzed in the original Woodbury and Novick article. The reliabilities required to disattenuate this matrix appropriately were .76, .82, .70, .64, and .74.

The particular population matrix used to generate the intercorrelated true ability and criterion parameters, given in Table 1, is partitioned. The last row and column of the matrix contains the true validity vector and the criterion self-correlation, unity. The larger partitioned area contains the intercorrelations between the latent or true abilities.

Table 1
The Assumed Population Matrix

	True Abilities						Criterion
Variable	1	2	3	4	5	6	7
1	1.00	.16	.47	.38	.46	.33	.50
2	.16	1.00	.22	.12	.21	.09	.17
3	.47	.22	1.00	.30	.53	.34	.51
4	.38	.12	.30	1.00	.49	.21	.36
5	.46	.21	.53	.49	1.00	.36	.50
6	.33	.09	.34	.21	.36	1.00	.20
7	.50	.17	.51	.36	.50	.20	1.00

Through the use of this matrix, 900 simulated cases were randomly sampled, each case having six true ability parameters, θ_{jk}, for $k = 1, 2, \ldots 6$ and one criterion parameter, y_j. The subscript j indexed the simulated cases where $j = 1, 2, \ldots N$.

In order to generate the item responses for the simulated cases for each item (ability) bank or test, the item parameters must be specified. The distinction between item bank and test depends on the use of the particular simulated case as described below. For convenience, each bank or test had the same parameter specifications. The parameters for the 100 items in each bank or test were specified as follows: In sequence, 20 items were assigned to each level of a_i, viz., .8, 1.2, 1.6, 2.0, and 2.4 that in turn contained 20 levels of b_i varying from -1.9 to 1.9 in increments of .2. The c_i were successively assigned the values from .03 to .27 in increments of .03, where both c_1 and c_{100} were accordingly .03. Again, to assure verisimilitude, these specifications were in accord with reasonable expectations for ability test items.

For 500 of these simulated cases, the generated item responses were scored as they would be for six 100-item conventional tests. Raw scores--merely the number of items answered correctly--were obtained for the 100-item test variables. The scores were then intercorrelated along with the criterion, Kuder-Richardson Formula 20 reliabilities were estimated, and a modified Woodbury and Novick solution was obtained.

Using the obtained solution, multidimensional tailored testing was then conducted with the 400 simulated cases remaining from the original 900. These cases were evenly divided into two samples, viz., Sample 1 and Sample 2. For each case, tailored testing proceeded by using each bank until the particular terminal standard error, $\tilde{\sigma}_{\varepsilon_k}$, was achieved. The tailored test scores, or ability estimates, were then weighted to obtain $\hat{y}_j$, the predicted or estimated criterion parameter. These estimates were then correlated with their corresponding and known true parameters in order to allow an assessment of the effectiveness of the multidimensional algorithm.

Results

The 100-item tests for the first 500 simulated cases were conventionally scored. Their reliabilities were computed by means of Kuder-Richardson Formula 20, and the tests were intercorrelated along with the criterion. The results are reported in Table 2. The off-diagonal elements of this matrix should resemble the off-diagonal elements of the assumed population matrix (as given in Table 1) from which the 500 simulated cases were sampled in order to allow the generation of item responses. The resemblance is unmistakable because both the sampling error for the 500 cases and the measurement error, as indicated by the high test reliabilities, were small.

Table 2

Obtained Reliabilities (Main Diagonal), Test Intercorrelations, and Validity Coefficients (Last Row and Column, Omitting the Main Diagonal), N=500

Variable	1	2	3	4	5	6	7
1	(.963)	.171	.510	.378	.422	.353	.512
2	.171	(.962)	.181	.122	.157	.083	.171
3	.510	.181	(.960)	.317	.516	.373	.469
4	.378	.122	.317	(.961)	.436	.221	.287
5	.422	.157	.516	.436	(.958)	.333	.487
6	.353	.083	.373	.221	.333	(.958)	.206
7	.512	.171	.469	.287	.487	.206	1.000

Using the data given in Table 2, the modified Woodbury and Novick procedure was applied. For this application, the initial testing times for the tests, t_k, were all set to unity. Accordingly, the diagonal matrix D_t was the identity matrix.

The pertinent results obtained in this application are summarized in Table 3. In the notation of this table, the squared construct validity coefficient, $\tilde{\rho}^2_{\hat{\theta}_k\theta_k}$, has been substituted for its identity, $\tilde{r}_{kk}$, the terminal reliability. This identity is readily established by merely correcting the reliability coefficient for attenuation in one of the parallel forms to obtain the validity coefficient. This coefficient is the correlation between the attenuated fallible parallel form and the disattenuated parallel form representing the errorless latent ability or pertinent psychological construct. Squaring this construct validity coefficient merely removes the radical, again yielding the reliability coefficient. This identity assumes that true score is a linear function of true ability.

In Table 3 terminal reliabilities ($\tilde{\rho}^2_{\hat{\theta}_k\theta_k}$), terminal standard errors ($\tilde{\sigma}_{\varepsilon_k}$), and the regressed estimate weights ($\tilde{b}_k$) are given for each bank at five levels of testing time (T^*) along with their associated maximum multiple correlations (R). It should be noted that the maximum multiple correlation increased with testing time and that these increases diminished in magnitude as testing time increased; the maximum multiple correlation as a function of testing time (T^*) eventually reached an asymptotic value beyond which further testing time (T^*) yielded no further return in validity (R). It should also be noted that relatively large increases in the terminal reliabilities were required for negligible increases in validity (R) as testing time (T^*) increased.

In this context, the terminal standard errors are completely determined by the terminal reliabilities. These are merely the square root of 1.0 minus the particular reliability. It is of interest to note that the banks with the higher terminal reliabilities, $\tilde{\rho}^2_{\hat{\theta}_k\theta_k}$, also have the larger regressed estimate weights, $\tilde{b}_k$. The ordering is perfect.

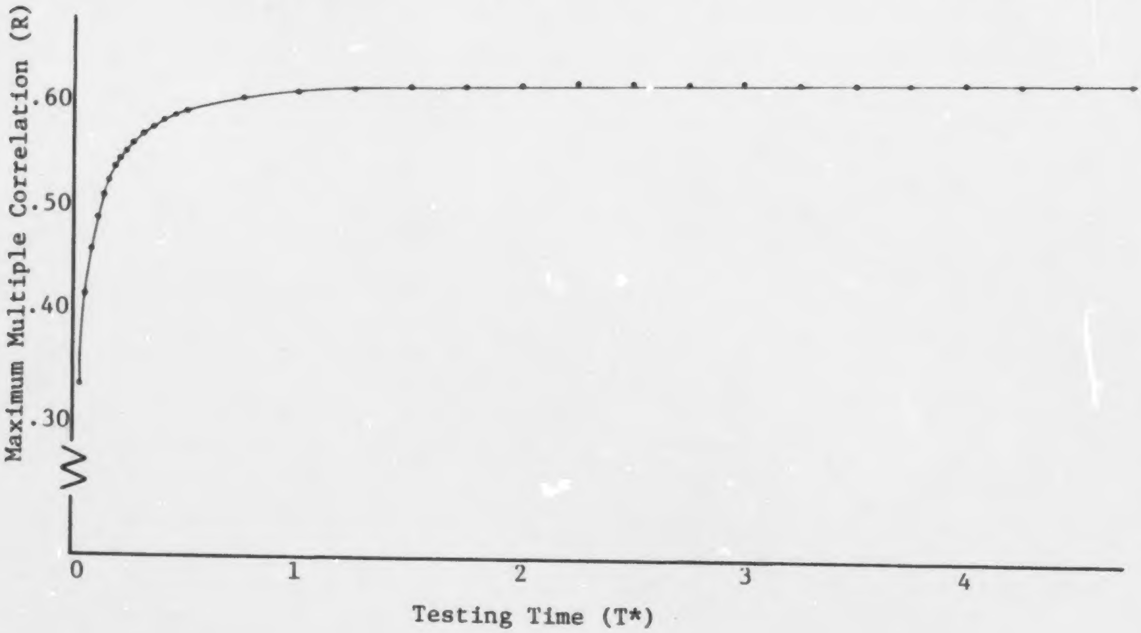
The asymptotic properties of the maximum multiple correlation (R) as a function of testing time (T^*) may be readily observed in Figure 1. Here this function is given for testing times (T^*) of zero through five. There is an abrupt rise in this function in the range of T^* of zero through one; thereafter, increases in the maximum multiple correlation tended to be negligible. The asymptote of the function is approximately .61. As a result, it was decided to use the modified solution at a T^* of 1, where the maximum multiple correlation was approximately .60. This solution yielded the terminal standard errors ($\tilde{\sigma}_{\varepsilon_k}$) and regressed estimate weights ($\tilde{b}_k$) for the six item

Table 3

Terminal Reliabilities ($\tilde{\rho}^2_{\hat{\theta}_k\theta_k}$), Terminal Standard Errors ($\tilde{\sigma}_{\varepsilon_k}$), and Regressed Estimate Weights ($\tilde{b}_k$) for Six Item (Ability) Banks at Five Levels of Testing Time (T^*), and the Associated Maximum Multiple Correlation (R)

	Item Bank (k)					
Testing Time (T^*)	1	2	3	4	5	6
1 (R=.598)						
$\tilde{\rho}^2_{\hat{\theta}_k\theta_k}$	.913	.440	.835	.000	.892	.000
$\tilde{\sigma}_{\varepsilon_k}$	.295	.748	.406	1.000	.329	1.000
$\tilde{b}_k$	.328	.052	.180	.000	.282	.000
2 (R=.607)						
$\tilde{\rho}^2_{\hat{\theta}_k\theta_k}$	.953	.690	.909	.000	.942	.000
$\tilde{\sigma}_{\varepsilon_k}$	.216	.557	.301	1.000	.241	1.000
$\tilde{b}_k$	.320	.049	.171	.000	.275	.000
3 (R=.610)						
$\tilde{\rho}^2_{\hat{\theta}_k\theta_k}$	.968	.786	.938	.000	.960	.000
$\tilde{\sigma}_{\varepsilon_k}$	.178	.463	.250	1.000	.199	1.000
$\tilde{b}_k$	.317	.048	.168	.000	.272	.000
4 (R=.612)						
$\tilde{\rho}^2_{\hat{\theta}_k\theta_k}$	.976	.835	.952	.000	.970	.000
$\tilde{\sigma}_{\varepsilon_k}$	.155	.406	.218	1.000	.173	1.000
$\tilde{b}_k$	.315	.047	.167	.000	.270	.000
5 (R=.613)						
$\tilde{\rho}^2_{\hat{\theta}_k\theta_k}$	.980	.867	.962	.000	.976	.000
$\tilde{\sigma}_{\varepsilon_k}$	.140	.365	.196	1.000	.156	1.000
$\tilde{b}_k$	.314	.046	.166	.000	.270	.000

Figure 1
The maximum multiple correlation (R) as a function of testing time (T*)



(ability) banks (k) that are indicated in columns 1, 2, and 3 in Table 4. If the square roots of the terminal reliabilities given in Table 3 for T^* equal to 1 are taken, they will represent forecasts that can be made relative to the $\rho_{\hat{\theta}_k \theta_k}$, or the constant validity of tailored test scores, where the criterion is the particular latent ability. These forecasts, the $\rho_{\hat{\theta}_k \theta_k}$, are given in row 5 of Table 4. The solution selected after the application of the modified procedure forecasts a cross-validity, $\tilde{\rho}_{yy}$, of .60 allowing, of course, for no shrinkage; this forecast is shown in row 4. It should be noted that the abilities measured by Banks 4 and 6 were not required in criterion performance. The terminal standard errors $\tilde{\sigma}_{\varepsilon_4}$ and $\tilde{\sigma}_{\varepsilon_6}$, both 1.00 for Banks 4 and 6, imply that $\tilde{\rho}_{\theta_4\theta_4}$ and $\tilde{\rho}_{\theta_6\theta_6}$ both equal zero. Thus, unidimensional tailored testing was unnecessary with respect to Banks 4 and 6.

Multi-bank tailored testing was then conducted using the 400 simulated cases remaining from the original 900. These cases were evenly divided into two samples, viz., Sample 1 and Sample 2. For each case, tailored testing proceeded by using each bank until the particular terminal standard error, $\tilde{\sigma}_{\varepsilon_k}$, was achieved. The four tailored test scores, the $\hat{\theta}_{jk}$, were then weighted to obtain $\hat{y}_j$, the predicted criterion score. The obtained correlations can be directly compared with the forecasts of theory provided in Table 4. These comparisons indicate that the multidimensional procedure performed very well

Table 4
Multidimensional Tailored Testing:
Forecasted and Obtained Results

Statistic		Item Bank (k)						
		1	2	3	4	5	6	
Terminal Standard Error	$\tilde{\sigma}_{\varepsilon_k}$	.30	.75	.41	1.00	.33	1.00	
Regressed Score Weight	$\tilde{b}_k$	.33	.05	.18	.00	.28	.00	
Forecasted Results:								
Validity	$\tilde{\rho}_{\hat{\theta}_k \theta_k}$	.95	.66	.91	.00	.94	.00	
Cross-Validity	$\tilde{\rho}_{yy}$							.60
Simulation Results:								
Sample 1								
Validity	$r_{\hat{\theta}_k \theta_k}$	.95	.70	.91	.00	.95	.00	
Average No. Items	$\bar{m}$	7.20	1.00	4.20	.00	6.10	.00	
Total No. Items	$\Sigma\bar{m}$							18.50
Cross-Validity	r_{yy}							.59
Sample 2								
Validity	$r_{\hat{\theta}_k \theta_k}$	.94	.68	.92	.00	.94	.00	
Average No. Items	$\bar{m}$	7.10	1.00	4.00	.00	6.00	.00	
Total No. Items	$\Sigma\bar{m}$							18.10
Cross-Validity	r_{yy}							.64

in terms of the reduction in the total of the average number of items required per examinee vis-a-vis the number of items typically used in conventional paper-and-pencil test batteries.

Discussion

When external criterion measures are available, the economy of multidimensional tailored testing derives from

1. low values for the terminal reliabilities,
2. a reduced number of measured abilities, and
3. an allocation of terminal reliabilities to minimize computer-interactive time.

Low values for the terminal reliabilities are possible because the maximum multiple correlation (validity) as a function of testing time typically approaches asymptotically high values quite rapidly. High values for the terminal reliabilities, requiring larger numbers of items, are necessary only after this function has approached the asymptote.

Abilities that are not valid or those sufficiently well measured through correlated abilities are not measured. Thus, correlated abilities pose no particular problem. A proper economy in items is observed even when abilities are correlated. If the items measuring different abilities vary systematically in average examinee response time, computer-interactive time is minimized through the use of the appropriate initial time matrix D_t in obtaining a modified solution. These initial testing times need only be proportional to attain a proper solution.

Whether a modified solution is found for conventional tests or tailored tests, it is equally applicable later in multidimensional tailored testing. The case in which the modified solution was found through conventional testing and later applied in multidimensional tailored testing was illustrated in this paper. The modified solution can also be found for tailored tests and then applied in multidimensional tailored testing.

Certain salient aspects of tailored testing will now be considered by means of the Owen (1969, 1975) algorithm. It is well known that the correlation coefficient completely determines the correlational surface. Thus, a function completely determined by the correlation coefficient will do likewise. For example, if the standard error of the estimate, $\sqrt{1 - \rho^2_{\hat{\theta}_k\theta_k}}$ (the error about the regression of θ_k on $\hat{\theta}_k$), is controlled by the appropriate termination of tailored testing, the slope of this regression, $\rho_{\hat{\theta}_k\theta_k}$, is also controlled. Since the correlational surface implies equality and symmetry of errors about both regression lines, the error about the regression of $\hat{\theta}_k$ on θ_k, which is more traditionally considered to be the standard error of measurement, is equal to the standard error of the estimate, $\sqrt{1 - \rho^2_{\hat{\theta}_k\theta_k}}$ (In this context, the traditional distinction between these standard errors, drawn from classical test theory, breaks down algebraically.)

This determination of the correlational surface also implies a marginal distribution of $\hat{\theta}_k$ that has a scaling identical to that of the marginal distribution of θ_k, true ability, where the mean is zero and the standard deviation is unity. To allow this feature of control over the correlational surfaces,

1. tailored testing with the Owen algorithm must begin with zero and unity for the prior estimates of ability and the standard error of the estimate, respectively; and
2. the scaling of a mean of zero and a standard deviation of unity for θ_k must have been employed when the item parameters were being estimated in large random samples from the population of interest.

Appropriate termination requires variable-length tailored tests to control the resulting correlational surfaces. Fixed-length tailored tests do not provide this control because the standard error of the estimate is, out of necessity, ignored. Evidence of the effectiveness of this control through variable-length tailored testing is provided in Table 4, as well as in several other studies (Urry, 1974, 1975, 1977).

In this context, fixed-length tailored tests clearly result in curvilinear regressions of estimated ability on the corresponding true ability because the standard error of the estimate is ignored. Equiprecision of measurement throughout the important range of each ability is lost when, as in fixed-length tailored testing, the standard error of the estimate is ignored. Terminating the variable-length tailored sequences after a specific value of the standard error of the estimate has been achieved guarantees equiprecision of measurement for the important range of each ability.[2]

Through the multidimensional algorithm, a conditional maximum multiple correlation or validity coefficient at a fixed testing time is sought. To attain the unconditional maximum validity coefficient, infinite testing time would be required. While this is clearly the solution, it has a severe practical drawback in requiring an infinite number of items. Fortunately, increases in testing time beyond some point yield negligible returns with respect to the validity coefficient. There is, then, a trade-off involved in which the appropriate testing time can be rigorously established in a decision theoretic framework. For example, a specific cost/benefit ratio can uniquely determine the appropriate testing time. In this context, the computer could be used to monitor (1) validities, (2) costs, and (3) benefits. Thus would be known on a continual basis (1) the validities of the procedures for personnel selection; (2) the dollar costs of obtaining pertinent test information; and (3) the dollar benefits in increased productivity accruing from the selection decisions.

Preliminary work in decision or utility theory has already been initiated at the U. S. Civil Service Commission. The findings indicate that the dollar benefits of personnel testing tend to be grossly underestimated by both practitioner and sponsor. A close examination of the value of personnel testing would afford a realistic reappraisal, which is much needed after the controversy surrounding personnel testing during the passing decade.

References

French, J. W. The validity of new tests for the performance of college students with high-level aptitude (Research Bulletin 63-7). Princeton, NJ: Educational Testing Service, 1963.

---

[2] In a Bayesian context, the standard error of the estimate is the proper term to use in determining equiprecision. The reciprocal square root of the information function is appropriate only in a maximum likelihood context. Error reduction is more rapid in the Bayesian context and occurs to a greater extent when incorrect answers are encountered. This can be deduced from Equation 3.7d provided by Owen (1975, p. 353). Greater efficiency in the Bayesian context can result in a correlation between the length of variable-length tailored tests and ability estimates because examinees of lower ability provide more incorrect answers. Hence, fewer itmes are required.

Horst, P. Determination of optimal test length to maximize the multiple correlation. Psychometrika, 1949, 14, 79-88.

Horst, P. Optimal test length for maximum differential prediction. Psychometrika, 1956, 21, 51-66.

Jensema, C. J. The validity of Bayesian tailored testing. Educational and Psychological Measurement, 1974, 34, 757-766.

Jensema, C. J. Bayesian tailored testing and the influence of item bank characteristics. Proceedings of the first conference on computerized adaptive testing (PS-75-6, U. S. Civil Service Commission,Personnel Research and Development Center). Washington DC: U.S. Government Printing Office, 1976. (Superintendent of Documents Stock No. 006-000-00940-9)

Owen, R. J. A Bayesian approach to tailored testing (Research Bulletin 69-92). Princeton, NJ: Educational Testing Service, 1969.

Owen, R. A. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of the American Statistical Association, 1975, 70, 351-356.

Taylor, C. W. A method of combining tests into a battery in such a fashion as to maximize the correlation with a given criterion for any fixed total time of testing. Unpublished master's thesis, University of Utah, 1939.

Urry, V. W. Computer-assisted testing: Calibration and evalution of the verbal ability bank (Technical Study 74-3). Washington, DC: U.S. Civil Service Commission, Personnel Research and Development Center, December 1974.

Urry, V. W. Computer-assisted testing with live examinees: A rendezvous with reality (Technical Research Note 75-3). Washington, DC: U.S. Civil Service Commission, Personnel Research and Development Center, January 1975.

Urry, V. W. A five-year quest: Is computerized adaptive testing feasible? Proceedings of the first conference on computerized adaptive testing (PS-75-6, U.S. Civil Service Commission, Personnel Research and Development Center). Washington, DC: U.S. Government Printing Office, 1976. (Superintendent of Documents Stock No. 006-000-00940-9)

Urry, V. W. Tailored testing: A successful application of latent trait theory. Journal of Educational Measurement, 1977, 14, 181-196.

Woodbury, M. A., & Novick, M. R. Maximizing the validity of a test battery as a function of relative test lengths for a fixed total testing time. Journal of Mathematical Psychology, 1968, 5, 242-259.

A MODEL FOR TESTING WITH MULTIDIMENSIONAL ITEMS

JAMES B. SYMPSON
UNIVERSITY OF MINNESOTA

Most of the recent research in adaptive testing has been based on latent trait theory (also known as item response theory and item characteristic-curve theory). This approach to test theory involves the use of mathematical models that relate item response probabilities to a person's status on one or more inferred, but unobservable, metric dimensions. While it will be several years before latent trait theory is widely applied in practical testing situations, some of the simpler models are already being utilized in commercially published tests (e.g., Woodcock, 1973); and a large-scale project at the U. S. Civil Service Commission is devoted to implementing computer-assisted, latent-trait-based testing (McKillip & Urry, 1976).

The mathematics involved in the use of latent trait models is a problem which has inhibited their practical application. This obstacle will be overcome with the appearance of simplified treatments of latent trait theory (e.g., Hambleton, Swaminathan, Cook, Eignor, & Gifford, 1977) and the dissemination of computer programs that can be used to estimate latent trait parameters (e.g., Kolakowski & Bock, 1973; Wood, Wingersky, & Lord, 1976; Urry, 1975). A second problem is that currently available latent trait models are inappropriate for certain practical testing situations. Most of the models do not apply when performance in the testing situation depends upon speed of response. New models suitable for such testing situations are needed. Also, most existing models assume that the probability of a correct response is a function of status on a *single* latent ability dimension. It is difficult to justify this assumption (either empirically or theoretically) with some types of psychological or educational tests.

Consider the test items shown in Figure 1. On the left are items for which a univariate latent trait model would probably be satisfactory. The item on the right, however, requires *both* verbal and numerical ability for its solution. If a person cannot read well enough to comprehend the problem or cannot complete the necessary arithmetic properly, a correct solution can only be achieved by guessing. This item taps at least two types of ability, and possibly a third (an ability to "integrate" verbal and quantitative information). For this type of item, a multivariate latent trait model is needed.

Figure 1
Unidimensional and Multidimensional Test Items

Examples of Unidimensional Test Items	Example of a Test Item That Involves Both Verbal and Numerical Abilities
Verbal Ability: Choose the word whose meaning is closest to that of the first word: HOT 1. COLD 2. WARM 3. MOIST 4. SOUR 5. HEAVY	If it takes 3 men 10 hours to dig a ditch 40 feet long, how long will it take these men to dig a similar ditch that is 50 feet long?
Numerical Ability: $2 + 7 + 4 + 8 = ?$	

Previously Developed Multivariate Latent Trait Models

Multivariate latent trait models have been developed by Rasch (1961), Samejima (1974), and Christofferson (1975). None of these models appears to have been utilized in applied testing situations. Each model has its unique limitations. The Rasch model assumes that all the item discrimination parameters are equal. Samejima's model assumes that the item response is a continuous function of latent trait status; it does not apply to dichotomously scored test items. Christofferson's model implicitly assumes that the distribution of latent trait ability parameters in the population of individuals to be tested is multivariate normal.

In addition to these individual limitations, all three of these models suffer from two further drawbacks. First, none deals systematically with the possibility that a correct response might occur as a result of lucky guessing. Second, all three of these models are fully compensatory in nature. That is, low status on one latent dimension can be completely offset by high status on another latent dimension. Consider again the item on the right in Figure 1. None of the previously developed multivariate latent trait models is appropriate for such an item. These models imply that a person with an arbitrarily low level of numerical ability could still have probability approaching 1.0 of passing the item if his/her verbal ability were high enough. What is needed is a multivariate latent trait model that allows for partial compensatory effects, but which does not lead to this incongruous result.

A New Multivariate Latent Trait Model

The following multivariate latent trait model is proposed for use with dichotomously scored multiple-choice items:

$$P_j^*(\underline{\Theta}) = c_j + (1. - c_j)\left(\prod_{l=1}^{m} [1. + exp[-1.7a_{jl}(\theta_l - b_{jl})]]^{-1}\right) \qquad [1]$$

$$= c_j + (1. - c_j)\left(\prod_{l=1}^{m} [P_j(\theta_l)]\right) = c_j + (1. - c_j)P_j(\underline{\Theta}) \ .$$

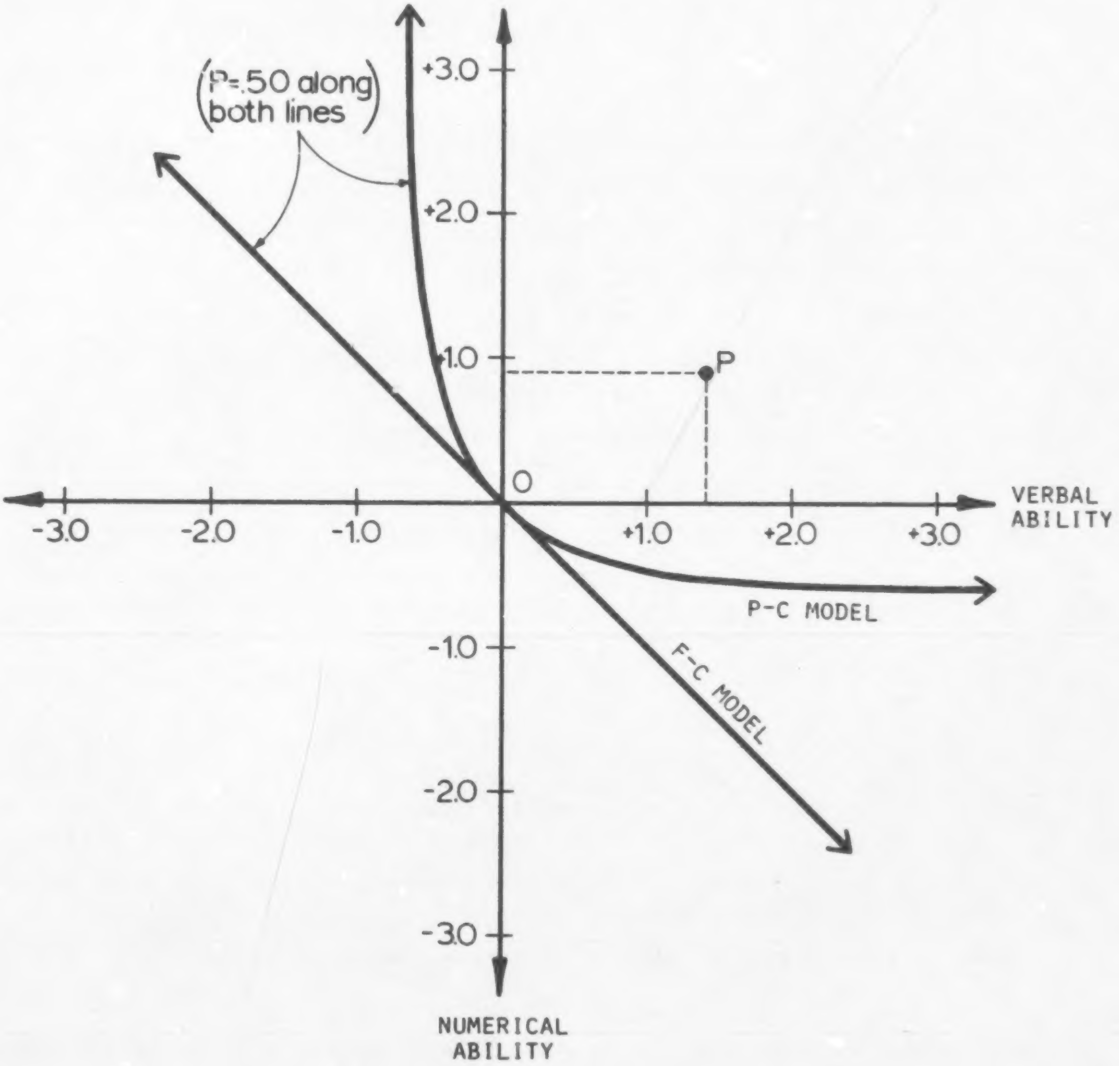
In this model, $P_j^*(\underline{\Theta})$ is the probability of occurrence of a correct response to item j by a person whose location in an m-dimensional latent space is described by the vector $\underline{\Theta}$. The parameters a_{jl} and b_{jl} index the discrimination and difficulty, respectively, of item j with respect to latent dimension l. The parameter θ_l is an element of $\underline{\Theta}$ and indexes the projection of a person onto dimension l. The constant -1.7 scales each latent dimension so that the quantity $P_j(\theta_l)$ equals approximately .8455 whenever $(\theta_l - b_{jl}) = 1/a_{jl}$. This choice of scaling constant causes the logistic function associated with each latent dimension to be approximately equivalent to a cumulative normal distribution function with mean equal to b_{jl} and standard deviation equal to the reciprocal of a_{jl}. The parameter c_j is the latent trait "guessing" parameter. It functions as a lower bound on $P_j^*(\underline{\Theta})$ as the probability of *knowing* the correct response,

$$P_j(\underline{\Theta}) = \prod_{l=1}^{m} [P_j(\theta_l)], \qquad [2]$$

approaches zero. For the balance of this discussion it will be assumed that c_j is equal to $1/A_j$, where A_j is the number of response alternatives available in item j. It is of interest to note that whenever m=1, the proposed multivariate model reduces to a univariate logistic model that has been investigated extensively by Birnbaum (1968).

In this multivariate model, each quantity $P_j(\theta_l)$ can be interpreted as the probability of a person with ability θ_l knowing the correct response to item j if he/she were possessed of infinite ability on each and every one of the m-1 latent dimensions other than dimension l. (Here it is assumed that a_{jl} is not precisely equal to zero on any of the latent dimensions. The value of b_{jl} is not unique if a_{jl} equals zero.) The lowest of the $P_j(\theta_l)$ values is an upper bound on $P_j(\underline{\Theta})$ that cannot be exceeded. However, below this upper bound, compensatory effects can occur.

Figure 2
Isoprobability Lines Under Two Different
Models in a Bivariate Latent Space



The logic of this model can be appreciated by considering the special case where one of the $P_j(\theta_l)$ values equals zero. This would occur if a person had an infinitely low projection onto one of the *m* latent dimensions. In this case, $P_j(\underline{\Theta})$ also equals zero. This is entirely appropriate for multidimensional items of the type presented in Figure 1. If a person has essentially no ability with regard to one of the skills required for correctly answering an item, then the probability that the person will answer the item correctly without guessing is necessarily equal to zero. No amount of ability with regard to other required skills will be sufficient to offset this fundamental deficit.

The difference between this new test model and fully compensatory multivariate models is illustrated in Figure 2. In this figure, verbal and numerical ability dimensions are represented by the horizontal and vertical axes, respectively. A person is represented in this space by a point with coordinates equal to the person's ability scores on the two dimensions. The point P in Figure 2 represents a person who is 1.4 standard deviations above the mean of zero on verbal ability and .9 standard deviations above the mean on quantitative ability.

Consider a test item, such as the one presented in Figure 1, requiring both verbal and quantitative ability for its solution. In particular, consider an item for which the probability of knowing the correct answer is equal to .50 when it is administered to a person located at the mean on both the verbal and numerical ability dimensions. (Such a person is located at the point O, where the horizontal and vertical axes intersect in Figure 2.) Shown in Figure 2 are the $P_j(\underline{\theta})=.50$ isoprobability lines for one such item, with $a_{j1}=a_{j2}$ and $b_{j1}=b_{j2}$, under a fully compensatory (F-C) test model and under the partially compensatory (P-C) model developed in this research.

The straight line passing through point O from upper-left to lower-right indicates that under the fully compensatory model, all people having abilities such that their verbal ability score equals minus the value of their numerical ability score will also have probability .50 of knowing the correct response to the item. This leads to the undesirable implication that a very low numerical ability score of -3.0 can be offset by a score of +3.0 on the verbal dimension.

In contrast, the curved line associated with the partially compensatory model shows that under this model a decrease in one ability can only be offset by a larger increase in the other ability. Outside of a relatively narrow ability range, the model becomes effectively non-compensatory. In this example, if a person's numerical ability score is below about -.55, the numerical score deficit cannot be offset by any empirically realizable increase in verbal ability. Conversely, if a person's verbal ability is low enough to interfere with comprehension of the question, increases in the numerical ability score will have an inconsequential effect on the probability of answering the question correctly. This partially compensatory model seems much more appropriate for multidimensional items of the type illustrated in Figure 1.

A Procedure for Estimating the Parameters of the New Model

It is not enough to propose a potentially useful model. In order to apply the model, it is necessary to develop a procedure for estimating the model's parameters. One possible method for estimating the parameters of the multivariate logistic model described above is comprised of three stages: (1) a regression stage, (2) a component-analysis stage, and (3) a parameter-fitting stage.

Regression Stage

In this stage of the estimation procedure, estimates of the nk quantities $P_j^*(\underline{\Theta})$ are first obtained for n persons who have responded to k dichotomously scored items. Then, the estimated $P_j^*(\underline{\Theta})$ values are corrected for guessing to obtain estimates of the nk values of $P_j(\underline{\Theta})$.

Assume that n people ($n>>k$) have been scored 0 or 1 on each of the k items to be calibrated. The binary item score variable can be designated X_{ij}, where $X_{ij}=1$ if person i gave a correct response to item j and $X_{ij}=0$ if an incorrect response was given. The observed value of X_{ij} may be predicted using multiple linear regression, with the values of $X_{ij'}$ from the k-1 items other than item j as predictor scores. Let the predicted value of X_{ij} be designated as $\hat{X}_{ij}$. As n becomes large, $\hat{X}_{ij}$ approaches $E[X_{ij}|v]$, where E is the expectation operator and v is the (k-1)-dimensional score vector associated with the performance of person i on the k-1 items other than item j. It can be readily shown that $E[X_{ij}|v] = E[P_j^*(\underline{\Theta})|v]$ in any population. Thus, a value of $\hat{X}_{ij}$ obtained by multiple linear regression can be viewed as an empirical Bayes point estimate of $P_j^*(\underline{\Theta})$ in a population that is characterized by the inter-item covariance matrix and vector of item means observed in the (large) sample of n persons.

The n-by-k matrix $\hat{\underline{X}}_{n,k}$ of predicted item responses can be generated with the matrix operation

$$\hat{\underline{X}}_{n,k} = \underline{X}_{n,k} - (\underline{X}_{n,k} - \underline{E}_{n,k})(\underline{C}_{k,k}^{-1}\underline{D}_{k,k}) , \qquad [3]$$

where $\underline{X}_{n,k}$ is the matrix of observed X_{ij} values;

$\underline{E}_{n,k}$ is a matrix whose rows are the k-dimensional row vector of observed item means repeated n times;

$\underline{C}_{k,k}$ is the inter-item covariance matrix; and

$\underline{D}_{k,k}$ is a diagonal matrix with diagonal elements equal to the reciprocals of the diagonal elements of $\underline{C}_{k,k}^{-1}$.

It has already been noted that n should be large in order to insure that $\hat{X}_{ij}$ is reasonably close to the population value of $E[X_{ij}|v]$. What about the magnitude of k? If k is small, the conditional variance of $P_j^*(\underline{\Theta})$, given the occurrence of a particular score vector v, will be relatively large. On the other hand, as k becomes large the conditional variance of $P_j^*(\underline{\Theta})$ associated with each possible v will decline. Thus, making k large will tend to reduce the average squared difference between $\hat{X}_{ij} \doteq E[X_{ij}|v]$ and $P_j^*(\underline{\Theta})$.

The regression procedure described above generates estimates of $P_j^*(\underline{\Theta})$ values. However, what is needed are estimates of $P_j(\underline{\Theta})$ values. From Equation 1, it can be seen that

$$P_j(\underline{\Theta}) = \frac{P_j^*(\underline{\Theta}) - c_j}{1 - c_j} \; . \qquad [4]$$

Thus, $\hat{X}_{ij}$ may be substituted for $P_j^*(\underline{\Theta})$ and the value $1/A_j = c_j$ inserted into this equation in order to obtain an estimate of $P_j(\underline{\Theta})$ for item j and person i. (With finite n, it is possible that $\hat{X}_{ij}$ may be less than c_j or greater than 1. If this occurs, either $(c_j + \delta)$ or $(1. - \delta)$, where δ is a small number such as .01, can be substituted for $P_j^*(\underline{\Theta})$ in Equation 4.) When this guessing correction is applied to all the $\hat{X}_{ij} = \hat{P}_j^*(\underline{\Theta})$ values, the n-by-k matrix $\underline{\hat{X}}_{n,k}$ is transformed into the matrix $\underline{\hat{P}}_{n,k}$, the elements of which estimate the $P_j(\underline{\Theta})$ values underlying the nk observed responses.

Component-Analysis Stage

In this stage of the estimation procedure, the nkm values of $P_j(\theta_l)$ underlying the observed responses are estimated. This is accomplished by factoring a rescaled cross-products matrix of $log[\hat{P}_j(\underline{\Theta})]$ values.

From Equation 2 it can be seen that

$$log[P_j(\underline{\Theta})] = log[\prod_{l=1}^{m} [P_j(\theta_l)]] \qquad [5]$$

$$= \sum_{l=1}^{m} log[P_j(\theta_l)] \; .$$

If an n-by-k matrix $\underline{L}_{n,k}$ were created by taking the log of each element in $\hat{\underline{P}}_{n,k}$, the m largest principal components of the k-by-k matrix $(\underline{L}'_{k,n}\underline{L}_{n,k})$ could be extracted. It would then be assured that the reproduced matrix $\hat{\underline{L}}_{n,k} = \underline{S}_{n,m}\underline{F}'_{m,k}$, where $\underline{S}_{n,m}$ contains component scores and $\underline{F}_{k,m}$ contains component "loadings," would be a least-squares fit to $\underline{L}_{n,k}$. Note that the (i,j)th element of $\hat{\underline{L}}_{n,k}$ is equal to

$$\sum_{l=1}^{m} [s_{il}f_{jl}] \, , \qquad [6]$$

where s_{il} is an element of $\underline{S}_{n,m}$ and
f_{jl} is an element of $\underline{F}_{k,m}$.

This suggests that the nkm quantities $s_{il}f_{jl}$ might be treated as approximations to $log[P_j(\theta_l)]$ values. Accordingly, approximations to the nkm values of $P_j(\theta_l)$ that underlie the observed item responses can be obtained by calculating the values of $antilog[s_{il}f_{jl}]$.

While this approach to estimating $P_j(\theta_l)$ values seems straightforward, there are some complicating considerations. First, a component analysis of the matrix $(\underline{L}'_{k,n}\underline{L}_{n,k})$ would be dominated by the most difficult items present in the analysis. This is because the lengths of vectors representing difficult items will be considerably greater than the lengths of vectors representing easy items. This fact becomes obvious when one considers that the base e log of $\hat{P}_j(\underline{\Theta}) = .05$ is -3.00 while the base e log of $\hat{P}_j(\underline{\Theta}) = .95$ is -.05. An easy item will have many small (negative) entries in its column of $\underline{L}_{n,k}$ while a difficult item will have many large (negative) entries in its column. To eliminate this undesirable effect of item difficulty, the principal components analysis can be applied to a rescaled cross-products matrix

$$\underline{W}^{-\frac{1}{2}}_{k,k}(\underline{L}'_{k,n}\;\underline{L}_{n,k})\underline{W}^{-\frac{1}{2}}_{k,k} \, , \qquad [7]$$

where $\underline{W}^{-\frac{1}{2}}_{k,k}$ is a diagonal matrix with diagonal elements equal to the reciprocal-square-roots of the diagonal elements in $(\underline{L}'_{k,n}\underline{L}'_{n,k})$. This rescaling brings all the item vectors to unit length so that each item has an equal influence on the component analysis. As will be seen below in Equation 9, this rescaling is taken into account when generating the estimated $P_j(\theta_l)$ values after the component analysis is completed.

Another important consideration is that an appropriate value for m, the number of latent dimensions, must be determined. In general, there is never more than a hypothesis about the value of m. However, if the items to be calibrated have been carefully written and carefully selected for inclusion in the analysis, the test developer's hypothesis should be reasonably accurate. Moreover, empirical evidence can always be used to help confirm this initial hypothesis. For example, the first m eigenvalues of the matrix defined in Equation 7 should account for a major portion of the matrix trace, while the last $(k-m)$ eigenvalues should be small and of nearly equal magnitude.

The third consideration to be dealt with is that any nonsingular rotation of the m principal components extracted will account for the input data equally well. That is, for any given value of m, there are an infinite number of sets of $s_{il}f_{jl}$ values that provide a least-squares approximation to $\underset{\sim}{L}_{n,k}$. One of these sets must be chosen. A wide variety of rotational procedures are currently available. Most of these attempt to satisfy one or more of Thurstone's simple-structure criteria (Thurstone, 1947, p. 335). These criteria involve specifications regarding the number and patterning of zero elements in the k-by-m reference-axis-structure matrix $\underset{\sim}{V}_{k,m}$ or, equivalently, the k-by-m primary-axis-pattern matrix $\underset{\sim}{F}_{k,m}$. (For a discussion of the distinction between reference axes and primary axes and the distinction between structure matrices and pattern matrices, see Rummel, 1970, pp. 396-408.)

While blind simple-structure rotations may be the best procedure for many applications of component or factor analysis, a more profitable approach can be followed when engaging in item calibration of the type discussed here. A specified simple-structure can be "built-in" when items are chosen for the analysis, and then the principal components can be rotated to a position that satisfies this specified simple-structure. To clarify this notion, first consider the implications of simple-structure in the context of the parameter estimation procedure being developed.

When an element of $\underset{\sim}{V}_{k,m}$ is equal to zero, the corresponding element of the primary-axis-pattern matrix $\underset{\sim}{F}_{k,m}$ will also be equal to zero. Conversely, whenever an element of $\underset{\sim}{F}_{k,m}$ is zero, the corresponding element of $\underset{\sim}{V}_{k,m}$ is also zero. As an element of $\underset{\sim}{F}_{k,m}$, say f_{jl}, approaches zero, all the products $s_{il}f_{jl}$ for item j must also approach zero. This causes all the $\hat{P}_j(\theta_l) = antilog[s_{il}f_{jl}]$ values for item j on latent trait l to approach 1. Earlier, it was noted that the quantity $P_j(\theta_l)$ can be interpreted as the probability of a person with ability θ_l knowing the correct response to item j if he/she were possessed of infinite ability on all the latent dimensions other than dimension l. This interpretation implies that any item for which all the $P_j(\theta_l)$ values are equal to 1 must be an item for which the probability of a correct response is not a function of ability on dimension l (or that the item requires so little of the ability that b_{jl} is effectively $-\infty$). Thus, specifying that an item should have a zero element in a given column of $\underline{F}_{k,m}$ (and, hence, $\underline{V}_{k,m}$) is equivalent to declaring that the item does not require any consequential amount of the type of ability with which the column's primary axis is associated.

The suggested approach to component rotation can be demonstrated by considering the calibration of k multivariate items of the type presented earlier in Figure 1. It was observed that such items require both verbal and numerical ability and, possibly, a third "integrative" ability. These two possibilities can be illustrated using two hypotheses regarding the number of latent dimensions, m=2 and m=3. Consider first the m=2 case. If several "pure" verbal items and several "pure" quantitative items (such as those illustrated on the left in Figure 1) were calibrated along with the multivariate items, zero values would be anticipated in the "verbal ability" column of $\underline{F}_{k,m}$ for the pure quantitative items and zero values would be anticipated in the "quantitative ability" column for the pure verbal items. The multivariate items would not be anticipated to produce zero elements in either column of $\underline{F}_{k,m}$. An appropriate rotation of the principal components should result in a reference-axis-structure matrix $\underline{V}_{k,m}$ that contains near zero values in the elements corresponding to anticipated zeros in $\underline{F}_{k,m}$.

Given these conditions, a procedure developed by Jöreskog (1965) may be utilized to determine the m-by-m transformation matrix $\underline{T}_{m,m}$ that will rotate the principal components into reference axes that minimize the sum of squares of the anticipated zero elements in $\underline{V}_{k,m}$. This procedure does not require the specification of the values of non-zero elements in $\underline{V}_{k,m}$. Once the transformation matrix $\underline{T}_{m,m}$ has been determined by Jöreskog's method, the matrix of primary-axis scores $\underline{S}_{n,m}$ can be generated with the matrix operation

$$\underline{S}_{n,m} = \underline{L}_{n,k}\, \underline{W}_{k,k}^{-\frac{1}{2}}\, \underline{A}_{k,m}\, \underline{\Lambda}_{m,m}^{-\frac{1}{2}}\, \underline{T}_{m,m}\, (\underline{T}'_{m,m}\, \underline{T}_{m,m})^{-1}\, \underline{D}_{m,m}^{-\frac{1}{2}}\,, \qquad [8]$$

where $\underline{L}_{n,k}$ and $\underline{W}_{k,k}^{-\frac{1}{2}}$ are as defined earlier;

$\underline{A}_{k,m}$ contains the first m eigenvectors of the rescaled cross-products matrix defined in Equation 7;

$\underline{\Lambda}_{m,m}^{-\frac{1}{2}}$ is a diagonal matrix with diagonal elements equal to the reciprocal-square-roots of the first m eigenvalues of the rescaled cross-products matrix;

$\underline{T}_{m,m}$ is the simple-structure transformation matrix; and

$\underline{D}_{m,m}^{-\frac{1}{2}}$ is a diagonal matrix with diagonal elements equal to the reciprocal-square-roots of the diagonal elements of $(\underline{T}'_{m,m}\underline{T}_{m,m})^{-1}$.

Next, the matrix of primary-axis-pattern coefficients $\underline{F}_{k,m}$ can be generated with the matrix operation

$$\underline{F}_{k,m} = \underline{W}_{k,k}^{\frac{1}{2}}\ \underline{A}_{k,m}\ \underline{\Lambda}_{m,m}^{\frac{1}{2}}\ \underline{T}_{m,m}\ \underline{D}_{m,m}^{\frac{1}{2}} \ . \qquad [9]$$

Finally, the desired values of $\hat{P}_j(\theta_l)$ can be generated by calculating $antilog[s_{il}f_{jl}]$ for all possible values of i, j, and l. Note that if $s_{il}f_{jl} \geq 0$, then $\hat{P}_j(\theta_l) \geq 1$. If this should occur, the value of $\hat{P}_j(\theta_l)$ can be set equal to $(1-\delta)$, where again δ is a small value such as .01.

This completes the component analysis stage for the hypothesis that m=2. What about the hypothesis that m=3? If an "ability to integrate verbal and quantitative information" does operate in these multivariate test items, it would be anticipated that pure quantitative items calibrated at the same time would have zero elements in both the "verbal ability" column and the "integrative ability" column of $\underline{F}_{k,m}$. Similarly, any pure verbal items would have zero elements in both the "quantitative ability" and "integrative ability" columns of $\underline{F}_{k,m}$. The multivariate items would not be anticipated to have any zero elements in $\underline{F}_{k,m}$. Given the pattern of zeros in $\underline{F}_{k,m}$ and, hence, $\underline{V}_{k,m}$ that this hypothesis implies, Jöreskog's procedure may again be applied in order to obtain a least-squares fit to the anticipated zero elements of $\underline{V}_{k,m}$. The primary-axis scores and the primary-axis-pattern coefficients can then be determined using Equations 8 and 9, and the $\hat{P}_j(\theta_l)$ values can be calculated.

The two examples discussed above were intended to demonstrate how Jöreskog's simple-structure rotation procedure can assist in the estimation of underlying $P_j(\theta_l)$ values. In order to apply this procedure, it is only necessary to select items for the analysis so that at least m (and preferably more) zero elements can be anticipated in each column of $\underline{F}_{k,m}$ and, hence, $\underline{V}_{k,m}$. While this approach seems particularly fruitful, other solutions to the component rotation problem could also be utilized.

It should be noted that if m=1 is hypothesized, the component-analysis stage of the estimation procedure can be omitted. When m=1, the $\hat{P}_j(\underline{\Theta})$ values obtained during the regression stage can be used as input data for the parameter-fitting stage.

Parameter-Fitting Stage

In this stage of the estimation procedure, an iterative least-squares algorithm is used to identify a set of item and person parameters that is consistent with the $\hat{P}_j(\theta_l)$ values obtained for each latent dimension ($l=1,2,\ldots,m$).

As can be seen in Equation 1,

$$P_j(\theta_l) = [1. + exp[-1.7a_{jl}(\theta_l - b_{jl})]]^{-1} . \qquad [10]$$

Rearranging this equation gives

$$[log_e[(1./P_j(\theta_l)) - 1.]/ - 1.7] = a_{jl}\theta_l - (a_{jl}b_{jl}) , \qquad [11]$$

a linear function of θ_l with slope coefficient equal to a_{jl} and additive constant equal to $-a_{jl}b_{jl}$. Let the function of $P_j(\theta_l)$ that appears on the left side of Equation 11 be referred to as $g[P_j(\theta_l)]$. Given n empirically determined values of $\hat{P}_j(\theta_l)$ derived from the earlier stages of the estimation procedure, n values of $g[\hat{P}_j(\theta_l)]$ can be calculated.

If the latent ability on dimension l of each of the n persons in the sample were known, bivariate linear regression could be utilized to determine the slope coefficient and additive constant that provide a least-squares fit to the $g[\hat{P}_j(\theta_l)]$ values. Let $\hat{P}_{ijl}$ be the value of $\hat{P}_j(\theta_l)$ and θ_{il} be the value of θ_l for person i. If the θ_l values have a mean of zero and unit variance, the following equations generate estimates of a_{jl} and b_{jl} that provide a least-squares fit to the empirically based $g[\hat{P}_j(\theta_l)]$ values:

$$\hat{a}_{jl} = \sum_{i=1}^{n}\left(\theta_{il}\ g[\hat{P}_{ijl}]\right)/n \qquad [12]$$

and

$$\hat{b}_{jl} = \sum_{i=1}^{n}\left(\theta_{il} - (g[\hat{P}_{ijl}]/a_{jl})\right)/n . \qquad [13]$$

Equation 13 could be further simplified by noting that

$$\sum_{i=1}^{n}\theta_{il} = 0 ,$$

since the mean of θ_l is assumed to be zero. However, in computer simulations it has been found that a modification of Equation 13,

$$\hat{b}_{jl} = \sum_{i=1}^{n}\left(w_{ijl}(\theta_{il} - (g[\hat{P}_{ijl}]/a_{jl}))\right)/\left(\sum_{i=1}^{n} w_{ijl}\right), \qquad [14]$$

leads to better parameter estimates. The quantity w_{ijl} is a weighting coefficient equal to

$$[\hat{P}_{ijl}(1. - \hat{P}_{ijl})]^{f(c_j)}, \qquad [15]$$

where the exponent $f(c_j)$ is a function of the item's "guessing" parameter and ranges from approximately 1/4 when c_j=.00 to approximately 1/8 when c_j=.30.

The effect of the weighting coefficient w_{ijl} is to make the estimate of b_{jl} most dependent on the responses of persons for whom $\hat{P}_{ijl}$ is near .50. As $\hat{P}_{ijl}$ approaches either zero or 1, indicating a great disparity between person i's ability and item j's difficulty, w_{ijl} approaches zero and person i is discounted as a source of information about item j's difficulty.

And what if a_{jl} and b_{jl} for all k items were known? In that case a least-squares estimate of θ_{il} is given by

$$\hat{\theta}_{il} = \sum_{j=1}^{k}\left((g[\hat{P}_{ijl}]/a_{jl}) + b_{jl}\right)/k. \qquad [16]$$

As before, a somewhat better estimate can be obtained using

$$\hat{\theta}_{il} = \sum_{j=1}^{k}\left(w_{ijl}((g[\hat{P}_{ijl}]/a_{jl}) + b_{jl})\right)/\left(\sum_{j=1}^{k} w_{ijl}\right). \qquad [17]$$

In this equation, items that are apparently very easy or very difficult for person i are given less weight.

Of course, neither the item parameters nor the ability parameters on dimension l are known. Nevertheless, estimates for all these parameters can be obtained by starting with rough approximations to the item parameters, estimating the ability parameters using Equation 17, using the obtained $\hat{\theta}_{il}$ values in Equations 12 and 14 to obtain improved item parameter estimates, re-estimating the ability parameters, and continuing in this manner until successive item parameter estimates differ by less than some small amount such as .001. While a mathematical proof that this iterative procedure will always achieve convergence cannot be offered here, in more than 100 computer simulations no convergence failures have been experienced. The procedure typically requires 4 or 5 iterations to achieve a convergence criterion of .001.

To begin the iterative process, it may be temporarily assumed that all the a_{jl} values are equal to 1. This assumption and the assumption that the θ_l values have a mean of zero lead to the following equation:

$$\sum_{i=1}^{n} g[\hat{P}_{ijl}] = \sum_{i=1}^{n} (a_{jl}\theta_{il} - (a_{jl}b_{jl})) = \sum_{i=1}^{n} (-b_{jl}) = -nb_{jl} . \qquad [18]$$

Thus, a rough approximation to b_{jl} is given by

$$\hat{b}_{jl} = -\left(\sum_{i=1}^{n} g[\hat{P}_{ijl}]\right)/n . \qquad [19]$$

Given these initial approximations to the item parameters, the first set of $\hat{\theta}_{il}$ values can be calculated and the iteration process set in motion.

The iterative parameter-fitting procedure described above can be applied first to the $\hat{P}_j(\theta_l)$ values for latent dimension 1 (l=1). After $\hat{a}_{j1}$, $\hat{b}_{j1}$, and $\hat{\theta}_{i1}$ values have been obtained and printed out by the computer, the same procedure can be applied to the $\hat{P}_j(\theta_l)$ values for latent dimension 2 (l=2). The process continues until parameter estimates for all m latent dimensions have been generated.

Preliminary Evaluation of the Procedure

At this time, the regression stage and the parameter-fitting stage of the estimation procedure have been programmed and are operational. Programming of the component-analysis stage is scheduled for the near future. Since this procedure uses a different statistical approach than other programs available for estimating latent trait parameters in the univariate (m=1) case, it seemed advisable to conduct some computer simulations with m=1. After all, if poor parameter estimates are obtained with m=1, there is little hope of obtaining good estimates when m>1.

Table 1 summarizes the results of 60 computer simulations with m=1. Thirty simulations were run with c_j=.00 for all items, and 30 simulations were run with c_j=.20 for all items. Each simulation in the c_j=.00 condition had a counterpart in the c_j=.20 condition. The same 30 sets of θ, a_j, and b_j values were used under each c_j condition. This allowed an evaluation of the effects of guessing that was not confounded by differences in the other parameters. To generate one set of parameters, 250 values of θ were drawn from a normally distributed population with $\mu=0.0$ and $\sigma=1.0$, 25 values of a_j were drawn from a normal population with $\mu=1.0$ and $\sigma=.3$, and 25 values of b_j were drawn from a normal population with $\mu=0.0$ and $\sigma=1.0$. For each possible pairing of person and item, the values of θ, a_j, b_j, and c_j were inserted into Equation 1 and the probability of a correct response was determined. This probability, in conjunction with a random number drawn from a uniform distribution on the interval (0,1), gave rise to a value of X_{ij}. The matrix $\underline{X}_{n,k}$ thus formed was then input to the regression stage of the parameter estimation procedure.

Before the regression calculations began, persons who correctly answered less than $(25c_j + 1)$ items and persons who correctly answered all the items were edited out of the analysis. Similarly, items answered correctly by fewer than $(250c_j + 10)$ persons and items answered correctly by more than 240 persons were edited out. Table 1 shows the average number of items and persons that remained after the editing process. The average number of iterations required to achieve a convergence criterion of .001 in the parameter-fitting stage is also shown.

Table 1
Results of 60 Computer Simulations*

Variable	$c_j = .00$	$c_j = .20$
Conditions		
Number of items	24.67	24.63
Number of persons	245.07	239.27
Number of iterations	3.67	4.87
Summary Statistics		
Mean		
$P_j(\theta)$	.482	.482
$\tilde{P}_j(\theta)$	.483	.485
Standard Deviation		
$P_j(\theta)$	.326	.326
$\tilde{P}_j(\theta)$	.335	.339
Regression of $\tilde{P}_j(\theta)$ on $P_j(\theta)$		
Slope	.982	.963
Intercept	.009	.020
Correlation	.954	.925

*Each table entry is the mean of 30 values.
(m=1 latent dimensions simulated)

After the parameter estimates $\hat{\theta}$, $\hat{a}_j$, and $\hat{b}_j$ were determined for a set of simulated data, the parameter estimates for each possible pairing of person and item were inserted into Equation 10 and a "reproduced" $P_j(\theta)$ value $[\tilde{P}_j(\theta)]$ was generated. Table 1 shows summary statistics and regression statistics comparing the $\tilde{P}_j(\theta)$ values with the true underlying $P_j(\theta)$ values. The quality of estimation, as shown by the data in Table 1, seems satisfactory. The estimated item and person parameters interrelate in a manner that mirrors the relationships which generated the binary data matrix $\underset{\sim}{X}_{n,k}$. Of course, the quality of the parameter estimates should improve as n and/or k are increased.

While these results for the m=1 case are encouraging, it should be kept in mind that when m>1 the estimation process will be subject to more error. It will probably be necessary to increase both n and k in order to get estimates that are as good as those obtained in these simulations. Hopefully, the required increases will not be too large.

References

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

Christoffersson, A. Factor analysis of dichotomized variables. Psychometrika, 1975, 40, 5-32.

Hambleton, R. K., Swaminathan, H., Cook, L. L., Eignor, E. R., & Gifford, J. A. Developments in latent trait theory: A review of models, technical issues, and applications. Paper presented at the joint meeting of the National Council on Measurement in Education and the American Educational Research Association, New York, April, 1977.

Jöreskog, K. G. On rotation to a specified simple structure (Research Bulletin 65-13). Princeton, NJ: Educational Testing Service, 1965.

Kolakowski, D., & Bock, R. D. LOGOG: Maximum likelihood item analysis and test scoring--logistic model for multiple item responses. Chicago: National Educational Resources, 1973.

McKillip, R. H., & Urry, V. W. Computer-assisted testing: An orderly transition from theory to practice. In C. L. Clark (Ed.), Proceedings of the First Conference on Computerized Adaptive Testing (U.S. Civil Service Commission, Personnel Research and Development Center, Professional Series 75-6). Washington, DC: U.S. Government Printing Office, 1976.

Rasch, G. On general laws and the meaning of measurement in psychology. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, 1961, 4, 321-334.

Rummel, R. J. Applied factor analysis. Evanston, IL: Northwestern University Press, 1970.

Samejima, F. Normal ogive model on the continuous response level in the multidimensional latent space. Psychometrika, 1974, 39, 111-121.

Thurstone, L. L. Multiple factor analysis. Chicago: University of Chicago Press, 1947.

Urry, V. W. Ancillary estimators for the item parameters of mental test models. Paper presented at the meeting of the American Psychological Association, Chicago, August,1975.

Wood, R. L., Wingersky, M. S., & Lord, F. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters (Research Memorandum 76-6). Princeton, NJ: Educational Testing Service, 1976.

Woodcock, R. W. Woodcock Reading Mastery Tests. Circle Pines, MN: American Guidance Service, 1973.

Acknowledgements

This project is supported by a research fellowship from National Computer Systems, Inc., Minneapolis, MN and the U. S. Office of Naval Research, contract N00014-76-C-0243, NR150-382, David J. Weiss, Principal Investigator.

The excellent work of Warren L. Cartwright as programmer for this research project is gratefully acknowledged.

DISCUSSION: SESSION 2

FREDERIC M. LORD
EDUCATIONAL TESTING SERVICE

Years ago I often wondered if it were possible to devise a nonparametric procedure that would be of practical use in mental test theory. I never did succeed in doing this. Now Cliff, Cudeck, and McCormick have succeeded; they have a very elegant procedure. It is elegant because it is so simple.

I have two observations to make about their procedure. First, the results that have been presented here are for situations in which there is no guessing. They have stated that their procedure worked best when there was no guessing. I would be interested to know what happens when there is guessing.

Second, Cliff, Cudeck, and McCormick state that their work is based on the Guttman scale model. They emphasize that the model does not fit their data. There is another way to say the same thing: Their procedure is not based on a model. This is not a criticism; on the contrary, it is an advantage. Nobody can now criticize their procedure on the grounds that it does not fit the model. Urry's model, my model, and Sympson's model are all open to severe criticism on these grounds; but there is no way to criticize Cliff's procedure. The only way to evaluate Cliff's procedure is to use real people and real test items; it cannot be evaluated with simulated data. Parallel forms of a Cliff-tailored test and parallel forms of another kind of tailored test should be obtained, and both should be administered to find out which testing procedure gives the least error of measurement. (Incidentally, I would be suspicious of using a product-moment reliability coefficient as a means of comparison in such a study.)

The Woodbury-Novick procedure on which Urry's paper is based depends very sensitively on the assumptions that (1) test reliability varies with test length according to the Spearman-Brown formula and (2) test validity varies with test length according to a similar familiar formula. Presumably, classical and conventional tests do behave in this way. It worries me when tailored testing is incorporated into this procedure. I do not believe that tailored tests, particularly when they are short tests, follow the Spearman-Brown formula or the other formula. Therefore, it is not clear to me that Urry is really minimizing testing time separately for each individual.

The Woodbury-Novick procedure is intended to maximize battery validity, a group statistic, for fixed testing time. In adaptive testing, something should be maximized for each individual rather than maximizing a group statistic. A situation might even be imagined in which there is only one individual rather than a group of people to be tested. Something should be maximized for this individual. I do not have any suggestion for dealing with this multidimensional tailored testing problem in a better way, however.

Sympson has pointed out, very appropriately, some disadvantages of certain multidimensional item response theories, and he has suggested a new model that avoids most of these difficulties. It appears to be a good model; upon a quick examination, I cannot see anything to criticize. Although it is not as general as it might be, this is true of every model, and particularly of multidimensional models.

Before returning to Sympson's paper, I want to make a suggestion. When there is a good probabilistic model for some data, it seems that a good procedure would be as follows: The frequency distribution of the observations should first be written down (a probabilistic model implies that this can be done, at least in principle); then, one of the standard statistical inference procedures, developed over the last fifty years, should be applied.

The typical problem is to use the observations to make inferences about model parameters or about some function of the model parameters. Two standard statistical inference procedures are maximum likelihood and Bayesian estimation. If the prior distribution for the parameters in a model are known, a Bayesian procedure should certainly be used; it will give better estimates than any other kind of procedure.

I do not wish to argue what kind of standard statistical inference procedure should be used, only that some standard procedure should be used. To illustrate, I will attempt to indicate what the maximum likelihood procedure would be for Sympson's model. The purpose of the following formulas is simply to illustrate that this can be done rather simply and directly without undue complications. Here

$i = 1, 2, \ldots, n$	examinees
$j = 1, 2, \ldots, k$	items
$\ell = 1, 2, \ldots, m$	ability dimensions
$P^*_j(\underset{\sim}{\theta})$	multidimensional item characteristic function
$P_j(\theta)$	unidimensional item characteristic function (not necessarily logistic)
u_{ij}	response of examinee i to item j (0 or 1)
$L \equiv L(U)$	joint distribution of u_{ij}
$I\{\ ,\ \}$	information

Equation 1 is Sympson's model; Equation 2 is its derivative:

$$P^*_{ij} \equiv P^*_j(\underset{\sim}{\theta}_i) \equiv c_j + (1 - c_j) \prod_{\ell=1}^{m} P_j(\theta_{\ell i}) \qquad [1]$$

$$\frac{\partial P^*_{ij}}{\partial \theta_{\ell i}} = (P^*_{ij} - c_j) \frac{\partial \log P_j(\theta_{\ell i})}{\partial \theta_{\ell i}} \qquad [2]$$

Equation 3 is the standard formula for the distribution of the observations in item response theory:

$$L(U) = \prod_{i=1}^{n} \prod_{j=1}^{k} P_{ij}^{*\,u_{ij}} Q_{ij}^{*\,1-u_{ij}} . \qquad [3]$$

The maximum likelihood estimates are the values of the parameters that maximize Equation 3. These are found by taking the derivative of the logarithm of the likelihood function in Equation 3 with respect to each of the parameters in turn and setting each derivative equal to zero. These equations are solved simultaneously to obtain the maximum likelihood estimates.

Equation 4 is the derivative of the log likelihood with respect to an ability parameter:

$$\frac{\partial \log\ L(U)}{\partial \theta_{\ell i}} = 0 = \sum_{j=1}^{k} \frac{u_{ij} - P_{ij}^{*}}{P_{ij}^{*} Q_{ij}^{*}} \frac{\partial P_{ij}^{*}}{\partial \theta_{\ell i}} . \qquad [4]$$

The derivative at the right is simply the quantity in Equation 2. For any given values of the responses and of the parameters, a computer could easily compute the quantity on the right. Equation 5 is the derivative with respect to an item difficulty:

$$\frac{\partial \log L(U)}{\partial b_{j\ell}} = 0 = \sum_{i=1}^{n} \frac{u_{ij} - P_{ij}^{*}}{P_{ij}^{*} Q_{ij}^{*}} \frac{\partial P_{ij}^{*}}{\partial b_{j\ell}} . \qquad [5]$$

Similar derivatives with respect to discriminating power, the *c* parameter, and other item parameters could be written.

If there are 1,000 people and two or three dimensions, this means there are 2,000 or 3,000 unknown parameters to solve for. This sounds appalling, but it is not. Whenever anyone scores 1,000 answer sheets, he/she is estimating 1,000 ability parameters. Furthermore, if number-right scores are used, each score is a good estimate of the ability of the person tested.

The standard procedure for solving maximum likelihood equations, such as Equations 4 and 5, is the Fisher scoring procedure, a modification of the well-known Newton-Raphson procedure for solving simultaneous equations. Only an information matrix is needed to carry out the Fisher scoring procedure. Elements of that matrix are given in Equations 6 and 7. Equation 6 shows an element of that part of the information matrix that relates to estimating examinee ability when item parameters are fixed:

$$I\{\theta_{\ell i}, \theta_{\ell' i'}\} \equiv \varepsilon \left(\frac{\partial \log L}{\partial \theta_{\ell i}} \frac{\partial \log L}{\partial \theta_{\ell' i'}} \right) \qquad [6]$$

$$= \begin{cases} \sum_{j=1}^{k} \frac{1}{P^*_{ij} Q^*_{ij}} \frac{\partial P^*_{ij}}{\partial \theta_{\ell i}} \frac{\partial P^*_{ij}}{\partial \theta_{\ell' i}} & \text{if } i = i' , \\ 0 \text{ otherwise} & . \end{cases}$$

Equation 7 gives an element of that part that relates to estimating item difficulty when ability parameters are fixed:

$$I\{b_{j\ell}, b_{j'\ell'}\} \equiv \varepsilon \left(\frac{\partial \log L}{\partial b_{j\ell}} \frac{\partial \log L}{\partial b_{j'\ell'}} \right) \qquad [7]$$

$$= \begin{cases} \sum_{i=1}^{n} \frac{1}{P^*_{ij} Q^*_{ij}} \frac{\partial P^*_{ij}}{\partial b_{j\ell}} \frac{\partial P^*_{ij}}{\partial b_{j\ell'}} & \text{if } j = j' , \\ 0 \text{ otherwise} & . \end{cases}$$

There would be similar formulas for estimating other item parameters.

In order to use the information matrix, it must be inverted. The question is, how can a matrix of order 2,000 or 3,000 be inverted? Fortunately, most of the elements of the information matrix are zero. In particular, if the item parameters are fixed, the information matrix needed for estimating examinee ability is virtually diagonal: It is a supermatrix of diagonal elements that are small matrices, 2 x 2 's or 3 x 3 's. In order to invert the entire 3,000 x 3,000 matrix, all that is necessary is to invert 1,000 2 x 2 or 3 x 3 matrices.

While estimating the ability parameters by the Fisher scoring method, the item parameters are fixed at some trial values. After estimates of the ability parameters are obtained, they are temporarily held fixed while Equation 7 and the corresponding Fisher matrix are used for estimating the item parameters. Then this process is repeated.

When used by itself, the Fisher scoring method is very efficient and effective; when it is used in this back-and-forth way, however, the total procedure takes considerable time. LOGIST is the program used at Educational Testing Service for doing this procedure unidimensionally. Typical LOGIST runs cost as low as $20. LOGIST does not do anything multidimensional, however, such as Sympson has proposed.

Although a maximum likelihood procedure could be used, the question arises whether or not there would be any advantage over Sympson's procedure. Sympson has done a very ingenious job; it is the sort of thing one would hope to do if one did not use maximum likelihood or some other standard statistical method. Nevertheless, Sympson's procedure has some drawbacks.

First, Sympson has estimated the response to a particular item from the responses to all the other items, using a multiple linear regression. The objection is that this regression is clearly and strongly non-linear; although multiple linear regression apparently gives him good results, the regression really should not be linear.

Second, Sympson has stated that when he corrects the probability of a correct answer for guessing, he sometimes obtains a negative value. Of course, a negative probability is impossible, so he replaces it with a small positive probability. This is not really harmful, but it is arbitrary. It is clearly not an optimal procedure.

Third, he is doing a factor analysis of a matrix consisting of logarithms or probabilities. It does not appear that this matrix satisfies the linear factor analytic model. He pointed out that the difficult items would get excessive weight in determining the results of the factor analysis; therefore, he rescaled the matrix to give all the items, in a certain sense, equal weight. But logically, some of the items may yield data that are more important than other items; the items should really not have equal weight. The problem is to know what weights should be assigned to which items.

Finally, Sympson replaces his data by an approximation obtained from the first two or three latent roots and vectors. This is a good approximation procedure, but I suspect that the residuals, which are discarded, are not completely devoid of information for purposes of making statistical inference. The main point I would like to make is this: When there is a probabilistic model describing the data, standard statistical methods should be used whenever possible.

SESSION 3

PSYCHOLOGICAL AND SUBGROUP EFFECTS

EFFECTS OF KNOWLEDGE OF RESULTS AND VARYING PROPORTION CORRECT ON ABILITY TEST PERFORMANCE AND PSYCHOLOGICAL VARIABLES

J. STEPHEN PRESTWOOD
UNIVERSITY OF MINNESOTA

STUDENT ATTITUDES TOWARD TAILORED TESTING

BILL R. KOCH AND
WAYNE M. PATIENCE
UNIVERSITY OF MISSOURI-COLUMBIA

REDUCTION OF TEST BIAS BY ADAPTIVE TESTING

STEVEN M. PINE
UNIVERSITY OF MINNESOTA

DISCUSSION

NANCY S. COLE
UNIVERSITY OF PITTSBURGH

SESSION 3: ABSTRACTS

EFFECTS OF KNOWLEDGE OF RESULTS AND VARYING PROPORTION CORRECT ON ABILITY TEST PERFORMANCE AND PSYCHOLOGICAL VARIABLES

J. STEPHEN PRESTWOOD

Testees were administered one of three conventional or one of three stradaptive vocabulary tests with or without knowledge of results (KR). The three tests of each type differed in difficulty, as assessed by the expected proportion of correct responses to the test items. Results indicated that the mean maximum-likelihood estimates of testee ability varied as a function of KR provision and test difficulty. Testees receiving KR scored highest on the most difficult test and lowest on the least difficult test; testees receiving no KR scored highest on the least difficult test and did most poorly on the most difficult test. Although testees perceived the differences in test difficulty, there were no effects on mean testee anxiety or motivation scores attributable to difficulty alone. Regardless of test difficulty, testees reacted very favorably to the provision of KR, which did increase mean testee motivation.

STUDENT ATTITUDES TOWARD TAILORED TESTING

BILL R. KOCH AND WAYNE M. PATIENCE

Two attitude scale questionnaires were administered to college students subsequent to tailored testing sessions to measure the interrelationships among perceived time pressure, test difficulty, test anxiety, and general test preference. Two instruments were used: (1) a Likert-type scale measuring attitudes on these dimensions and (2) a three-part attitude survey measuring attitudes toward five different test modalities, effects of computer familiarity on these attitudes, and preference for a black-on-white compared to a white-on-black CRT display screen. The relationships between student attitudes toward tailored testing and such variables as ability levels, perceived difficulty of the test, self-paced versus paced testing, and motivation are analyzed; and attitudes toward tailored tests and traditional tests are compared.

REDUCTION OF TEST BIAS BY ADAPTIVE TESTING

STEVEN M. PINE

Previous research on the effects of adaptive testing on the test-taking motivation of minority testees is replicated,and the relative effectiveness of computerized adaptive tests designed to minimize test bias is studied. Latent-trait calibrated item parameters were used to form two standard paper-and-pencil tests and two computerized adaptive tests, with one of the paper-and-pencil tests and one of the computerized tests designed to minimize test bias. Each of 250 high school students from two racial subgroups took one paper-and-pencil and one computerized test. In addition, half of the subjects were given feedback after each item was administered. The results indicated that the influence of feedback was more effective when given by computer than when provided in a standard test booklet. It was also found that the bias-reducing adaptive testing strategy, while maintaining a reasonable level of test reliability, was more effective at reducing test bias than was the paper-and-pencil test.

EFFECTS OF KNOWLEDGE OF RESULTS AND VARYING PROPORTION CORRECT ON ABILITY TEST PERFORMANCE AND PSYCHOLOGICAL VARIABLES

J. STEPHEN PRESTWOOD
UNIVERSITY OF MINNESOTA

Cronbach (1970, p. 35) has described ability tests as "those that seek to measure the maximum performance of the subject." It has long been recognized that for testees to achieve maximum levels of performance, they must be sufficiently motivated. The provision of immediate feedback or "knowledge of results" (KR) during testing has often been looked upon as one method for increasing testee motivation. On-line computerized testing has made the provision of KR a relatively simple matter. The ease with which KR can be administered is an added advantage of computerized adaptive testing which lies beyond the purely psychometric benefits of such procedures.

To study the effects and possible benefits of computer-administered KR, Betz and Weiss (1976a,b) administered multiple-choice tests of verbal ability to college undergraduates at the University of Minnesota. The tests were administered either with or without KR after each item response. Their data showed higher testee performance, as measured by maximum likelihood ability estimates, for students in the KR condition. Perceptions of test difficulty were more accurate for testees receiving KR; and these students also exhibited higher levels of motivation, as assessed by post-test measurements. Betz and Weiss' data also indicated that students' reactions to the provision of KR became more favorable as the proportion of positive feedback increased.

Because KR increased testee performance and motivation, and because testees reacted more favorably to the provision of KR as the proportion of positive feedback increased, an analysis of the joint effects of KR provision and the proportion of positive feedback (test difficulty) was initiated.

Method

Procedure

Subjects. Participating in this study were 561 undergraduate students enrolled in an introductory psychology course at the University of Minnesota in the fall of 1975. All subjects were volunteers who received points towards their final course grade for participation in the experiment. Students were sequentially assigned to experimental conditions.

Test administration. The students were tested at individual cathode-ray terminals (CRTs) connected to a Hewlett-Packard computer system. Instructional screens explaining the operation of the CRTs preceded the actual testing, and a proctor was present in the testing room to provide assistance in the operation of equipment. During the test, items were presented on the CRT screen; students responded by typing in a number corresponding to the chosen alternative for each of the 50 five-alternative multiple-choice vocabulary items.

Independent variables. A three-way factorial design was employed in the study. One factor was immediate knowledge of results (KR). Testees in the KR condition were informed by the computer immediately after a response whether it was correct or incorrect. After an incorrect response they were also told which of the alternatives was correct. Testees in the no-KR condition received no feedback. Another factor was ability-test strategy. Testees received either a conventional peaked ability test or a stradaptive ability test (Weiss, 1973). The third factor was test difficulty or proportion of positive feedback.

Three conventional tests and three stradaptive test-administration procedures were designed so that testees, on the average, would answer approximately 40%, 60%, or 80% of the test items correctly. Level of difficulty--high, medium, or low--was inversely related to the proportion of positive feedback a testee received, whether that feedback was explicit as in the KR condition or subjective as in the no-KR condition.

Items were chosen for the three peaked conventional tests on the basis of their normal ogive difficulty and discrimination parameters so that students would, on the average, be expected to answer 40%, 60%, or 80% of the questions correctly. The stradaptive tests were designed by constructing a stratified item pool with items grouped into nine non-overlapping difficulty strata. The items within each stratum were then arranged in decreasing order of discrimination.

The stradaptive branching routine normally branches to a different stratum depending on whether or not the preceding item was answered correctly (Weiss, 1973). For this study the procedure was modified so that branching to a more difficult stratum occurred whenever the current overall proportion-correct score for an individual was greater than a target value; and branching to a less difficult stratum occurred whenever the current proportion-correct score for an individual fell below the target value. The specific target values employed were determined by simulations and were chosen so that the actual final proportion-correct scores would be approximately .40, .60, or .80, as appropriate for the high-, medium-, and low-difficulty tests.

Dependent variables. Both the ability-test performance and the psychological reactions of the testees were of interest in the present study. Testee performance was measured by maximum-likelihood scores computed for each testee by solving the likelihood equation for the three-parameter logistic model of Birnbaum (1968, p. 459). Proportion-correct scores were also computed in order to ascertain the accuracy with which the target test difficulties were obtained. These latter scores, however, were not used as ability measures per se, since the stradaptive test-administration program was designed to yield an arbitrary proportion-correct for each testee, and testees received different items. Furthermore, differences in proportion-correct scores between testees were predetermined, to a degree, by the construction of the three different conventional tests which were administered to them.

The psychological reactions of testees to the task were determined using the responses of testees to rating-scale items administered following the test. The four scales constructed from these items measured a testee's perception of the test's difficulty; his/her level of anxiety during testing; his/her motivation to do well on the test; and, for each testee in the KR conditions, his/her reactions to the provision of explicit feedback.

The scales in this experiment were factor analytically derived and differed slightly from those employed in Betz and Weiss (1976b). The alternatives on the scale items were weighted so that items received equal weighting on the scale and so that increasing scores on the four scales corresponded to the following: increasing motivation, increasing anxiety, increasingly positive reactions to the provision of KR, and perceptions of increasing difficulty.

Results

Ability-Test Data

Proportion-correct measures. Table 1 shows the mean and standard deviation of the proportion-correct measures for each experimental condition. The table shows that, on the average, the tests in each condition achieved the appropriate target proportion-correct with a good degree of accuracy.

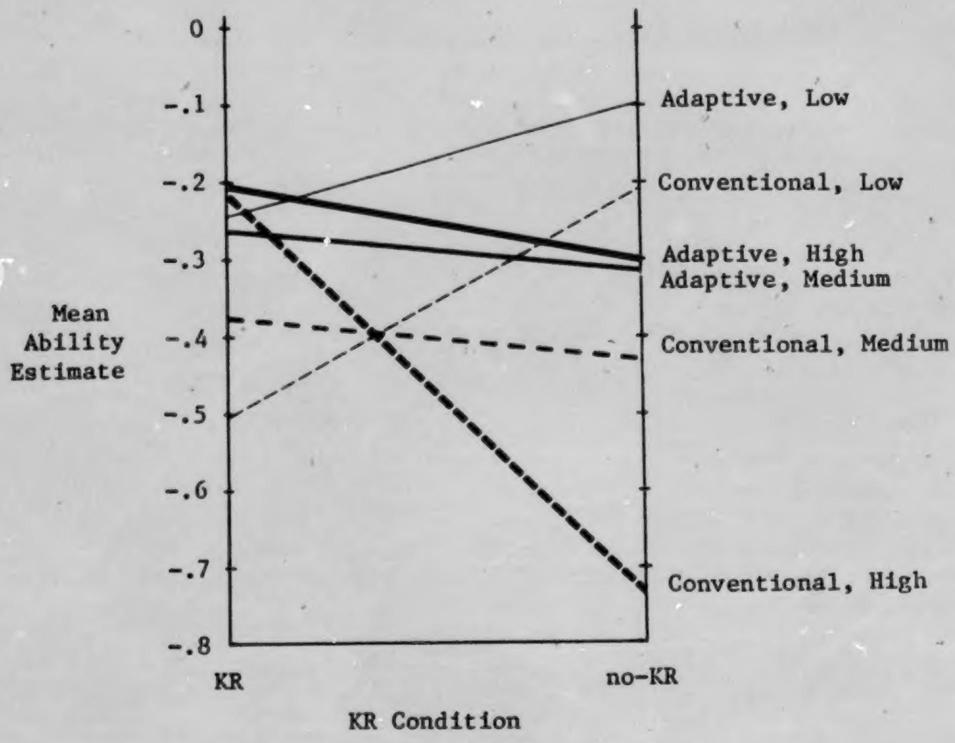
Table 1
Means and Standard Deviations of Proportion-Correct Scores for Conventional and Stradaptive Tests With and Without KR at Three Levels of Difficulty

Experimental Condition	KR *N*	KR Mean	KR S.D.	No-KR *N*	No-KR Mean	No-KR S.D.
Conventional Test						
Low Difficulty	48	.783	.106	45	.808	.103
Medium Difficulty	47	.608	.147	49	.592	.141
High Difficulty	46	.451	.188	46	.364	.153
Stradaptive Test						
Low Difficulty	44	.828	.064	45	.824	.046
Medium Difficulty	49	.617	.031	47	.610	.041
High Difficulty	49	.434	.103	46	.417	.076

The largest discrepancy (.051) was that between the target value of .400 and the actual value of .451 for the low-difficulty conventional test administered with feedback.

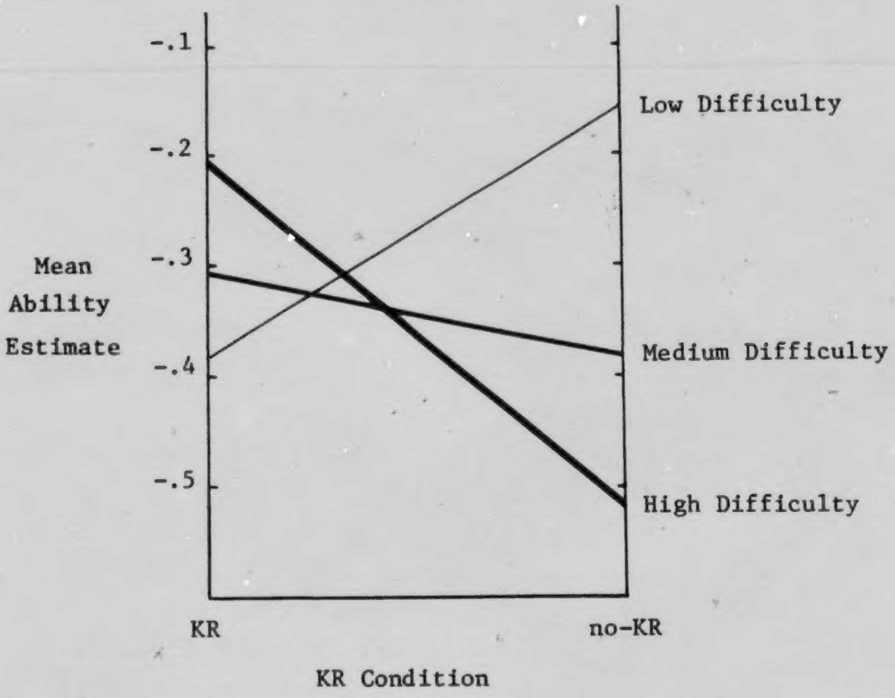
Maximum-likelihood ability estimates. Figure 1 shows the mean maximum-likelihood ability estimate for each of the 12 experimental conditions. The means for the KR conditions are plotted on the left vertical axis; the means for the conditions without KR are plotted on the right vertical axis. A three-way analysis of variance on the maximum-likelihood ability estimates showed a significant main effect for test strategy, $F(1, 549) = 3.984$, $p<.05$. The mean ability estimate for testees administered the stradaptive tests (-.239) was significantly higher than the mean for testees taking the conventional tests (-.415). There was also a marginally significant two-way interaction between the KR-provision and the test-difficulty factors, $F(2, 549) = 2.905$, $p \approx .054$. This interaction is shown graphically in Figure 2.

Figure 1
Mean Maximum-Likelihood Ability Estimates for
Adaptive and Conventional Tests of
Differing Difficulty Levels, by KR Condition



In Figure 2 the stradaptive and conventional test means have been combined. Interestingly, the effects of test difficulty on test performance were directionally opposite depending on whether or not KR was provided. When KR was provided, the mean testee ability estimate was highest on the most difficult tests (-.209) and lowest on the relatively easy tests (-.375). The mean ability estimates for testees in the no-KR conditions were highest on the easy tests (-.156) and lowest on the difficult tests (-.523).

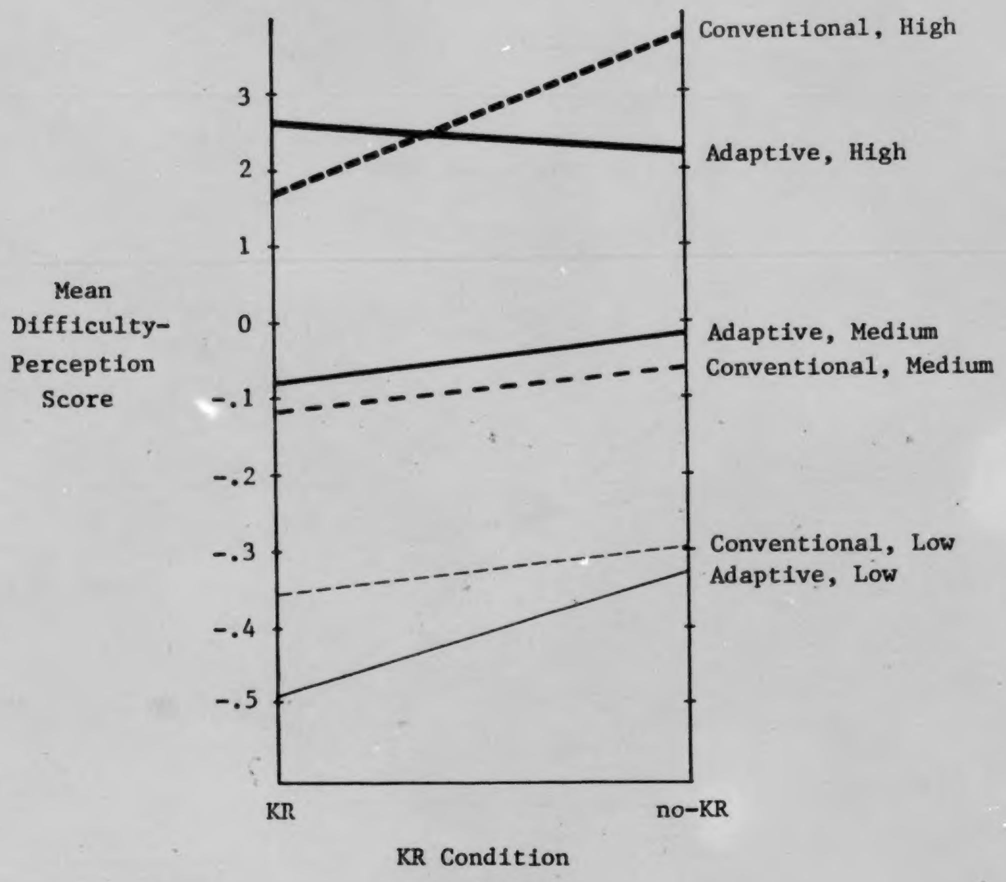
Figure 2
Mean Maximum-Likelihood Ability Estimates As
a Function of KR Condition and Test Difficulty



Psychological Reactions

Figure 3 shows the mean difficulty-perception scores as a function of experimental condition. A three-way analysis of variance assessing the effects of KR provision, test strategy, and test difficulty on the difficulty-perception data showed the expected main effect of test difficulty, $F(2, 549) = 163.243$, $p<.01$, and an additional main effect for the KR variable $F(1, 549) = 8.334$, $p<.01$. Tests administered without KR were perceived as significantly more difficult than tests administered with KR. There was also a marginally significant three-way interaction, $F(2, 549) = 2.810$, $p\approx.059$, due to the fact that the adaptive tests administered without KR differed less in perceived difficulty than did the adaptive tests administered with KR.

Figure 3
Mean Difficulty-Perception Scores for
Adaptive and Conventional Tests of
Differing Levels, by KR Condition

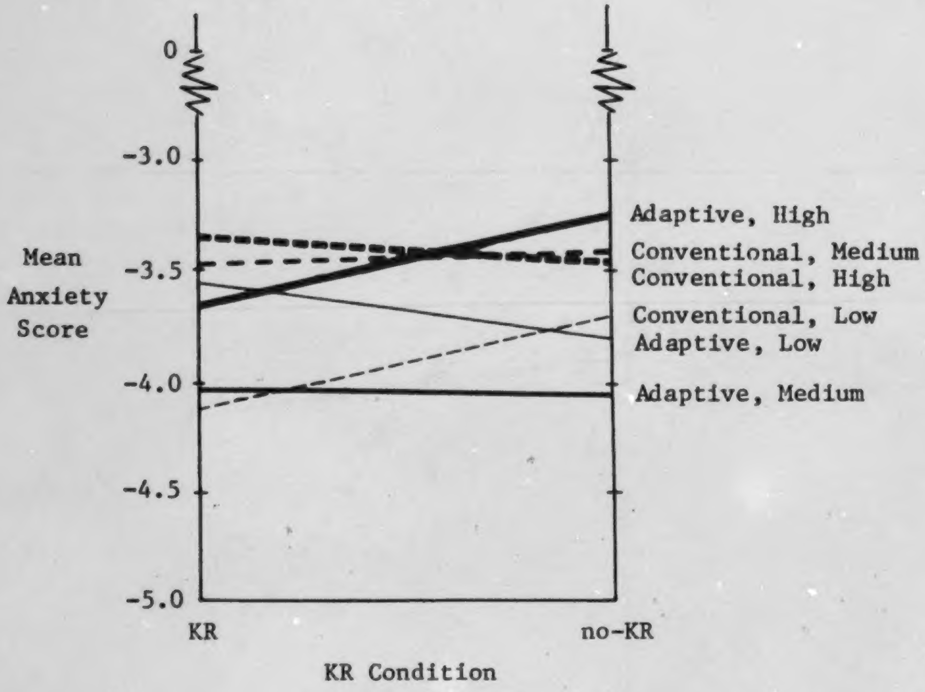


Mean anxiety scores are plotted as a function of experimental condition in Figure 4. An analysis of variance of these scores showed no effects of KR provision, test strategy, or test difficulty on mean level of anxiety.

Figure 5 shows the mean motivation scores for each of the 12 experimental conditions. A three-way analysis of variance of the motivation-scale data indicated a significant main effect for the KR factor, $F(1, 549) = 5.098$, $p<.05$. The mean motivation score for testees receiving KR (.096) was significantly higher than that for testees not receiving KR (-.110).

Figure 6 shows the mean KR-reaction score plotted as a function of test strategy for the high-, medium-, and low-difficulty tests. A two-way analysis

Figure 4
Mean Anxiety Scores for
Adaptive and Conventional Tests of
Differing Levels, by KR Condition



of variance assessing the effects of test strategy and test difficulty on testee reactions to KR showed no significant effects of experimental condition upon mean KR-reaction scores.

The endorsement frequencies of response-options on the KR-reaction questions are shown in Table 2. Of the 283 testees receiving immediate feedback, 87% felt that feedback made the test much more interesting; 86% felt that feedback did not interfere with their ability to concentrate on the test; 76% reported that feedback did not make them nervous; and 81% were very interested in knowing whether their answers were right or wrong. Ninety-two percent indicated that they liked getting feedback.

Table 3 shows the Pearson-product-moment correlation coefficients between the four psychological reactions scales. Those correlations involving the KR-reaction scale were based on the 283 students in the KR conditions. The other correlations were based on all 561 testees. As perceived difficulty increased, anxiety scores increased and motivation scores decreased. Reported anxiety was positively but not highly correlated with motivation scores. Students receiving KR reacted more favorably to KR as the perceived difficulty of the tests decreased.

Figure 5
Mean Motivation Scores for Adaptive and
Conventional Tests of Differing Difficulty Levels, by KR Condition

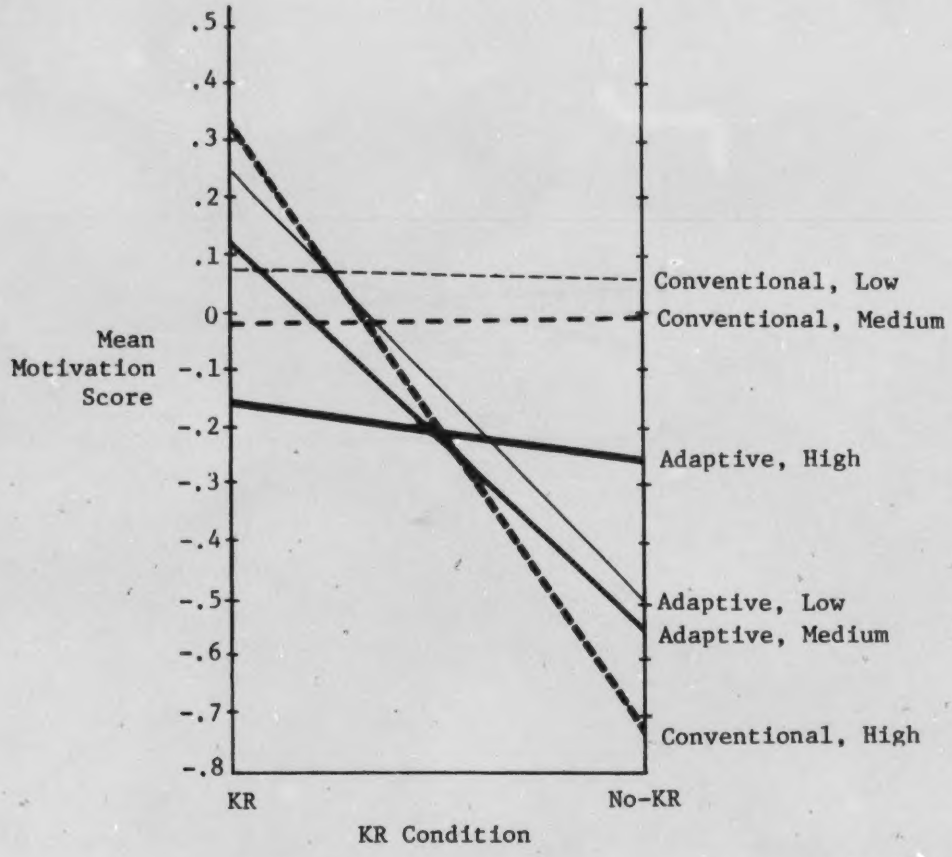


Figure 6
Mean KR-Reaction Scores for Adaptive and
Conventional Tests of Differing Difficulty Levels

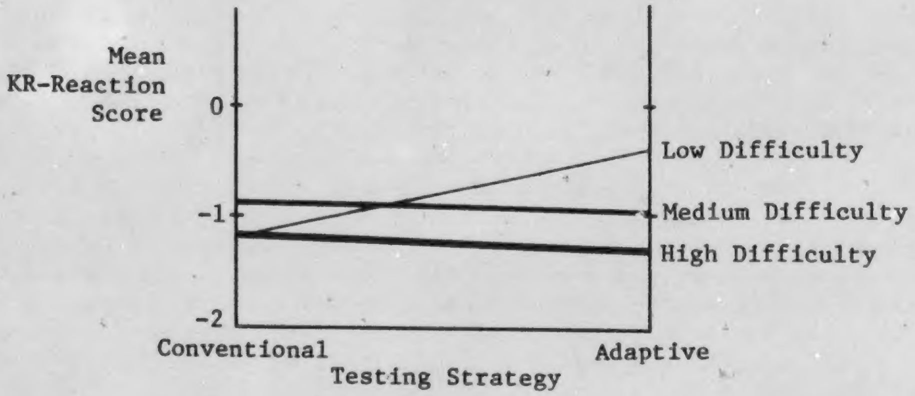


Table 2
Relative Frequencies of Response-Option Endorsement for KR-Reaction Items

Item	Endorsement Proportion
Did getting feedback on this test make it more interesting or less interesting?	
1. Much more interesting	.87
2. Somewhat more interesting	.11
3. Didn't make any difference	.00
4. Somewhat less interesting	.01
5. Much less interesting	.01
Did receiving feedback after each question interfere with your ability to concentrate on the test?	
1. No, not at all	.86
2. Yes, somewhat	.12
3. Yes, moderately so	.04
4. Yes, very much so	.02
Did getting feedback after each question make you nervous?	
1. No, not at all	.76
2. Yes, somewhat	.22
3. Yes, moderately so	.01
4. Yes, very much so	.01
Were you interested in knowing whether your answers were right or wrong?	
1. I was very interested	.81
2. I was moderately interested	.14
3. I was somewhat interested	.04
4. I didn't care at all	.00
How do you feel about getting feedback?	
1. I'd rather not know whether my answers were right or wrong	.07
2. I really don't care whether I get feedback or not	.01
3. I liked getting the feedback	.92

Table 3
Intercorrelations of Psychological-Reactions Scales

	Scale			
Scale	DIF	ANX	MOT	KR
DIF	-	.21**	-.17**	-.15*
ANX	.21**	-	.13**	-.16*
MOT	-.17**	.13**	-	.25**
KR	-.15*	-.16*	.25**	-

* Significant at the .05 level.
** Significant at the .01 level.

Discussion

The present data do not replicate the Betz and Weiss (1976a) finding that KR increased the average ability estimate of testees. Nor did these data find that anxiety was higher on the adaptive test--a finding reported by Betz and Weiss (1976b). The data in the present study showed higher ability estimates for students taking the adaptive test. Betz and Weiss (1976a) reported a similar effect, but only for a group of relatively lower ability students. There is agreement, however, between the present study and the Betz and Weiss (1976b) study, that the average motivation is higher for testees receiving KR. The increase in motivation accompanying the provision of KR may be partially due to the fact that testees receiving KR perceived the tests as being less difficult, on the average, than did testees not receiving KR. Also, it was noted that the scores on the difficulty-perception scale had a significant, although modest, negative correlation with motivation scores.

The marginally significant, but highly provocative, interaction of test difficulty and KR in the analysis of ability estimates indicates that the provision of KR may affect the performance of testees differentially, depending on the difficulty of the task. Although students reacted very favorably to KR regardless of the proportion of positive feedback, it would seem to be important that the effects of KR's provision *on test performance* be carefully investigated before it is provided under new sets of conditions.

References

Betz, N. E., & Weiss, D. J. Effects of immediate knowledge of results and adaptive testing on ability-test performance (Research Report 76-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1976. (NTIS No. ADA027147)(a)

Betz, N. E., & Weiss, D. J. Psychological effects of immediate knowledge of results and adaptive testing (Research Report 76-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1976. (NTIS No. AD027170) (b)

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), Statistical theories of mental Test Scores. Reading, MA: Addison-Wesley, 1968, 532-552.

Cronbach, L. J. Essentials of psychological testing (3rd ed.). New York: Harper & Row, 1970.

Weiss, D. J. The stratified adaptive computerized test (Research Report 73-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1973. (NTIS No. AD768376)

Acknowledgments

This research was supported by contract No. N00014-76-C-0243, NR150-382 with the Personnel and Training Research Programs of the Office of Naval Research, David J. Weiss Principal Investigator.

STUDENT ATTITUDES TOWARD TAILORED TESTING

BILL R. KOCH
WAYNE M. PATIENCE
UNIVERSITY OF MISSOURI--COLUMBIA

As tailored testing procedures gain in popularity and are frequently applied in testing situations, it becomes increasingly important to determine the psychological aspects of the tailored testing environment which may be introducing error into the test scores (Weiss, 1975). On the surface, favorable attitudes toward tailored testing would be expected due to such inherent characteristics as self-pacing of progress through the test, reduction in test length (both time and number of items), and matching of item difficulties to individual ability levels. All of these features contrast with traditional multiple-choice tests.

Some concern has been voiced, however, regarding a possible increase in the examinee's anxiety level during a test that involves the novelty of interaction with the computer. In addition, a frequent complaint from students that test questions on traditional tests did not cover the material they knew (and hence did not measure their true abilities) might be amplified in tailored tests, where the total number of items administered may average only 10 rather than the 50 typical of traditional tests (English, Reckase, & Patience, 1977).

Unfortunately, little of the published tailored testing research has reported on measures of examinees' attitudes toward the procedures. One study (Hedl, O'Neil, & Hansen, 1973), which attempted to determine the attitudinal effects of computerized intelligence testing, employed a five-item scale to measure anxiety level. It also employed a brief attitude scale to measure preference for the computerized test as compared to examiner-administered tests. The results indicated significantly higher anxiety levels and significantly less favorable attitudes for the computerized test than for the regular examiner-administered test. However, these results were probably due to an artifact of the study, since the computerized test (non-tailored) was full length, and examinees were required to complete all items, even if they reached the test ceiling by failing 10 consecutive items. Examinees who had such a failure experience scored significantly higher in anxiety level than the persons who did not reach the ceiling on the test. Thus the overall findings of this study were distorted.

A subsequent study (Lushene, O'Neil, & Dunn, 1974) attempted to measure the congruent validity of the MMPI as administered by a computer compared to the traditional booklet form; an anxiety level measure was incorporated as well. The results of the anxiety data indicated that the computer test initially produced higher anxiety levels in the examinees than the booklet form, but that this anxiety quickly dissipated once the computer session was underway; and there were no differences in anxiety levels at the end of the testing.

In this same area of personality assessment, there has been substantial research conducted using computers to administer personality instruments in an attempt to standardize the testing environment and eliminate the biases that may be induced with human examiners. Several studies hypothesized that examinees would respond more openly and honestly to highly personal or threatening items presented by the computer rather than by the human test administrator, but no significant findings have emerged in that direction (Resmovic, 1977). The proposed explanation for these results was that the studies failed to utilize the full interactive capabilities of the computer, using it instead as simply a presentation device.

Recently, Betz and Weiss (1976) conducted a comprehensive study which examined such attitude dimensions as level of motivation, anxiety level, perceived test difficulty, and immediate feedback of results on vocabulary tests. One important finding was that motivation was greater for low-ability examinees taking computerized adaptive (tailored) tests than for conventional tests administered on the computer. Another result was that significantly more anxiety was reported on the adaptive test than on the conventional test, even though both were computer administered. Also, students were able to perceive the difficulty of the test fairly well, although they were less consistently able to do so for the adaptive test. Finally, the immediate feedback feature was received very favorably by the examinees.

The attitude research on tailored testing has been conducted in relation to ability or personality testing rather than achievement testing. The difference, of course, is that in ability and personality tests, the examinees are typically asked to do their best or to respond honestly; but they have no clear incentive to do so. Achievement tests, on the other hand, are routinely used to assign course grades or for classification or placement decisions. One purpose of the present research, therefore, was to provide an indication of the attitudinal effects of tailored testing in the achievement test setting. Findings similar to previous research were expected in regard to perceived difficulty and computer interface effects; but differences were expected in such areas as anxiety and motivation levels, since in some cases achievement test results were used for course grades.

Instrumentation

Two separate attitude questionnaires were administered during the course of the present studies. The first instrument was a Likert-type scale consisting of four statements which measured attitudes toward tailored tests on the dimensions of (1) time pressure, (2) perceived test difficulty, (3) test anxiety, and (4) general test preference. The questionnaire was administered subsequent to examinees' tailored testing sessions in three separate research studies. In each case the attitude measures were secondary to the overall thrust of the research.

The second instrument was a three-part attitude survey which was administered during the course of the fourth study. The initial section of the questionnaire consisted of four items. Each item asked the examinee to rank five different test modalities along the dimensions of (1) perceived difficulty, (2) time pressure, (3) anxiety or stress level, and (4) overall preference. For example, Item 1 read as follows:

1. Assume that you have a test coming up in some course that you are taking. For the 5 types of tests below, please rank them into an order of difficulty. The type of test that is most difficult for you should be ranked 5, while the easiest test should be ranked 1.
    _____ true-false test (paper-and-pencil)
    _____ essay test
    _____ multiple-choice test (paper-and-pencil)
    _____ oral examination
    _____ computer-administered multiple-choice test

The items for time pressure, anxiety, and overall preference were nearly identical in format to Item 1. The order of the five types of tests, however, was varied to reduce the likelihood of a fixed response set during the ranking procedure. The design of this section of the questionnaire was based on Coombs' (1964) unfolding theory: the items attempted to determine existence of a latent attribute underlying preferences for the five types of tests (an ordered metric scale) for each of the four attitude dimensions. The unfolding technique makes minimal assumptions in finding the order of the stimuli, as well as the size of the distances between them.

The second part of the questionnaire consisted of just three items in which the examinees' prior experience with computers was measured. The items read as follows:

Are you at all familiar with computers?
____ Yes _____ No
Have you ever punched computer cards at a keypunch machine before?
____ Yes _____ No
Have you ever interacted with a computer by means of a terminal before?
____ Yes _____ No

In the subsequent analysis of the data, the response scores to this section of the questionnaire were used as covariates with the four unfolding items in order to determine the effects of computer familiarity on the attitudes expressed.

The third part of the attitude survey consisted of a scale of six Likert-type items in which each statement was followed by the alternative responses from which the examinee was to choose. The purpose of this section of the survey was to determine the examinees's relative preference for a black-on-white compared to a white-on-black cathode-ray-terminal (CRT) display screen; however, the statements did not make any specific reference to that purpose. This particular section of the survey was administered twice to each examinee: once after the tailored test in the black-on-white mode and once after the white-on-black test mode. Listed below are some examples of the statements:

1. The viewing screen was uncomfortable on my eyes.
   strongly agree agree neutral disagree strongly disagree
5. It was very easy to read the words and questions on the screen.
   strongly disagree disagree neutral agree strongly agree
6. Reading the questions on the screen was not much different from reading them on a regular test on paper.
   strongly agree agree neutral disagree strongly disagree

The response choices were weighted from 1 to 5 in the usual Likert fashion for scoring.

It should be noted that in neither of the two separate attitude questionnaires was the examinee responding anonymously, since the individual's student identification number was recorded at the time of the questionnaire administration.

Tailored Testing Research Designs

The attitude data reported in this paper were collected as supplementary parts of four different experimental designs. The four-item Likert questionnaire was administered during the course of the first three studies, all of which compared computerized tailored testing to traditional paper-and-pencil achievement tests; the second attitude questionnaire was administered during the fourth study.

The initial study investigated the reliability and validity of tailored testing compared to traditional achievement tests (Reckase, 1977). The test content covered the statistics and measurement portion of an introductory course in educational measurement and evaluation at the University of Missouri. The study employed a test-retest design (test sessions one week apart) with the attitude questionnaire being administered after the second session. Although the tailored test did not count toward the students' grades in the course, the students did receive extra credit for their participation.

The second study (English et al., 1977) was primarily concerned with measuring differences in levels of achievement for students taking tailored tests compared to those taking traditional paper-and-pencil tests. The examinees were enrolled in an introductory course in educational measurement and evaluation at the University of Missouri. Again, a test-retest design (test sessions three weeks apart) was employed, with the attitude questionnaire being given to the examinees after their second session. In this case, however, the results of the tailored tests were used in the assignment of course grades, thus providing an evaluation of the procedure under motivated circumstances.

The third study investigated the effects on performance in achievement tests of paced versus self-paced scheduling of the time of tailored test administration. In addition, tailored test performance was compared with traditional paper-and-pencil tests. The self-paced groups could take the tests whenever they wished and as often as they liked until satisfied with their grades. In contrast, the paced and traditional groups were scheduled to take the test at a specific time and could take it only once. Again, the tests counted toward course grades, and the attitude questionnaire was administered subsequent to the second test session.

The final experiment was essentially a pilot study concerned with applying the one- and three-parameter logistic models to tailored achievement tests. The purpose was to check out programs and procedures in preparation for subsequent live-testing studies. A counterbalanced experimental design was employed in which each examinee had two test sessions approximately one week apart. If the examinee took the test for the first session on the black-on-white CRT, then the second session would be on the white-on-black CRT, and vice versa. Lighting conditions in the test room were held constant, as were the CRT screen brightness and contrast controls.

The three-part attitude questionnaire was first administered after the examinee's initial test session. The third part only (dealing with the CRT screen display mode preferences) was re-administered following the second session, yielding attitude data for each display screen format.

The achievement test itself dealt with the evaluation techniques section of an introductory measurement and evaluation course at the University of Missouri. All students had previously just completed the traditional paper-and-pencil test for this section of the course. Therefore, although extra credit was given for participation in the study, the tailored tests did not count toward course grades.

Attitude Research Designs

In order to address the issue of examinees' attitudes toward tailored testing under unmotivated as compared to motivated conditions, a comparison was made between (1) the examinees' attitude responses to the questionnaire in the first study, where the tailored test did *not* count for course grades and (2) the responses on the same questionnaire in the second study, where the test did count. Obviously, this design was vulnerable to internal validity considerations, since the two groups were not based on random assignment, but were established depending on which semester the students took the course. The tailored test in both cases, however, covered identical course material and the physical testing conditions were equivalent (same test room and CRT terminal). Also, in both studies the examinees' participation was voluntary. This should have limited any possible differential selection effects of the first study compared to the second. Prior to analysis, the responses of an equal number of examinees from each group were randomly selected for comparison.

In addition to the analysis discussed above, each of the first three studies was subjected to a series of correlation analyses to measure the relationships between attitudes and such variables as ability levels, number of test items administered, and time spent taking the tailored test. Proportions of responses for each of the alternatives to the attitude items were also calculated to permit simple descriptive comparisons.

The second attitude questionnaire, which was administered as part of the fourth study, had three main research purposes. First, the responses to the Coombs' unfolding items were tabulated in order to determine the sets of individual preference rank orderings (called *I*-scales) of the five test types for each of the four attitude dimensions. These *I*-scales were then manipulated according to the unfolding technique to see if one dominant scale for each dimension (*J*-scale or joint continuum) could be recovered, upon which most of the respective *I*-scales would fit.

The second phase of the research was to convert the resulting *J*-scales from ordered metric scales into approximate interval scales, so that numerical values could be assigned to each of the positions along the *J*-scales. Upon completion, the scale values for each of the four attitude dimensions were related to ability levels, number of items administered, time spent taking the test, and prior computer experience, by means of multivariate analysis of variance procedures.

Finally, the last part of the questionnaire also used the multivariate analysis of variance technique to measure differences in responses to the six Likert-type items. The analysis compared attitudes toward the white-on-black CRT screen display to those toward the black-on-white CRT screen display. ADDS Consul 980 CRT terminals, which have both capabilities, were used in this part of the study.

Results

Motivated vs. Nonmotivated Groups

As can be seen in Table 1, the multivariate analysis of variance performed on the four attitude items for the motivated group compared to the unmotivated group yielded a statistically significant difference, approximate F (4,69) = 6.249, $p < .001$. The cell means indicate that the differences between the groups were observed in relationship to the time pressure and anxiety dimensions. The subsequent one-way analyses of variance to compare the groups on these dimensions yielded F (1,72) = 9.88, $p < .01$ for time pressure and F (1,72) = 4.11, $p < .05$ for anxiety. No significant differences were found regarding perceived test difficulty or overall preference for the tailored test compared to traditional paper-and-pencil tests.

These results indicate that the examinees in the motivated setting (where the tailored test counted toward their course grades) felt that the tailored test had less time pressure than the unmotivated group did. This was accompanied by higher anxiety levels during the tailored test for the motivated group.

Table 1
Means and MANOVA for Attitudes in
Motivated vs. Unmotivated Settings

	Cell Means	
Variable	Motivated	Unmotivated
Time Pressure	2.73	2.22
Difficulty	1.97	1.76
Preference	1.95	1.78
Anxiety	1.68	2.00

Univariate Analyses of Variance

Source	*SS*	*df*	*MS*	*F*
Time Pressure				
Mot./Unmot.	4.88	1	4.88	9.88**
Error	35.57	72	0.49	
Anxiety				
Mot./Unmot.	1.95	1	1.95	4.11*
Error	34.11	72	0.47	

$^{**}p < .01$;
$^{*}p < .05$

Likert Attitude Items

The response data to the four attitude items have been summarized in Table 2 for the three studies in which the first questionnaire was administered. Again, it is interesting to note the differences in response percentages to the alternatives for each item for Study 1 (unmotivated) compared to Studies 2 and 3 (motivated). In addition, it is possible to observe the overall attitude responses of the examinees toward tailored tests as compared to traditional tests on the four items of the questionnaire.

For example, regarding the dimension of time pressure, the majority of unmotivated examinees felt equal time pressure for both types of tests, while the majority of motivated examinees felt less time pressure on the tailored test. In terms of perceived difficulty, 70% of the unmotivated examinees found the tailored test more difficult, while the motivated examinees tended to find the tests equally difficult. Opinion appears to be about equally divided for all the examinees regarding overall test preference, although there is a tendency toward preference for the tailored test. Finally, although the motivated examinees tended to find that the tailored testing aroused as much as or more anxiety than traditional tests, the opposite was true for unmotivated examinees.

The results of the correlational analyses were inconclusive; individual correlation coefficients varied substantially for a given pairing of variables across tests. Only a few consistent correlations emerged, such as the findings that high ability examinees tended to find the tailored test easy and that examinees receiving higher numbers of tailored test items tended to find the tailored test more difficult.

Table 2
Likert Attitude Items and Response Data

Item	Response Percentages Study 1 (*N*=64)	Study 2* (*N*=41)	Study 3* (*N*=85)
1. Compared to multiple choice tests, the tailored test has			
a. more time pressure	19%	7%	11%
b. less time pressure	26%	78%	66%
c. about equal time pressure	55%	15%	23%
2. Compared to traditional multiple choice tests, the tailored test is			
a. easier	23%	25%	8%
b. harder	70%	31%	27%
c. about as difficult	7%	44%	65%
3. As compared to the traditional multiple choice test,			
a. I would rather take the tailored test	42%	44%	57%
b. I would rather take the traditional test	25%	44%	29%
c. I prefer both equally well	33%	12%	14%
4. Taking the test on the computer makes me			
a. more anxious than a traditional test	30%	42%	42%
b. less anxious than a traditional test	45%	12%	21%
c. about equally as anxious as the traditional test	25%	46%	37%

* In Studies 2 and 3 the tailored test counted toward the course grade, but not in Study 1.

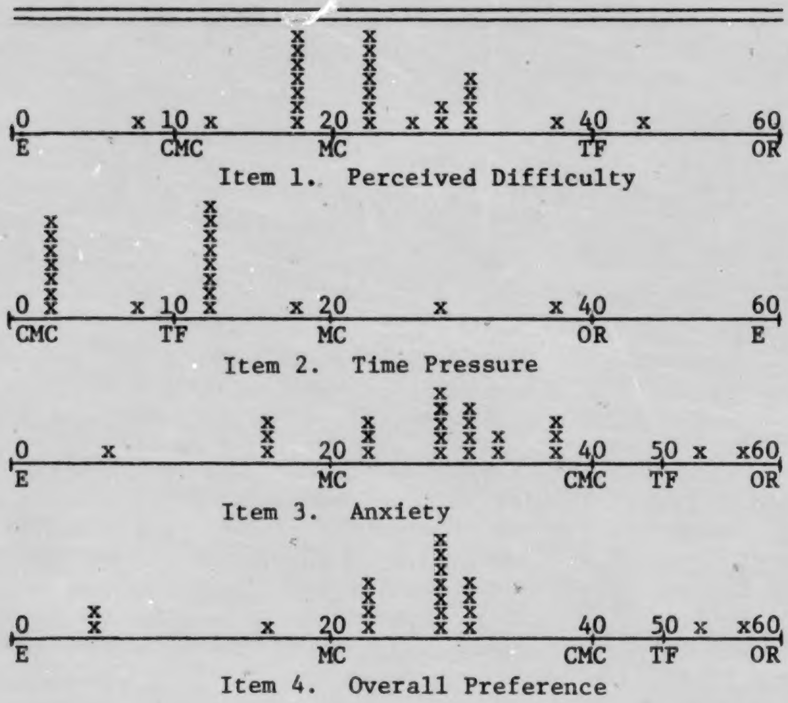
Unfolding Items

The results of the analysis of the responses to the four Coombs unfolding items from the second attitude questionnaire are presented in Figure 1. First, it can be seen that only three dominant *J*-scales or joint continua were found for the four attitude dimensions, since the anxiety continuum is identical to the overall preference continuum. The *J*-scales have been converted from ordered metric scales into approximations of interval scales. This shows the order in which the five test types fall along each continuum, as well as the relative distances in terms of preference between the five tests on the scales (Coombs, 1964). In addition, the frequencies of the *I*-scales are indicated with "X's" above each *J*-scale. Each *I*-scale denotes the position of an examinee on the scale in regard to preference for each of the five tests; the closer a person's position to a test, the more it is preferred.

For example, there were two dominant *I*-scales for the Perceived Difficulty *J*-scale. A total of 14 examinees found multiple-choice tests (MC) to be least difficult and computer multiple-choice tests (CMC) to be slightly more difficult. Essay (E) and true-false (TF) were even more difficult, and oral exams (OR) were the most difficult.

Figure 1

J-Scales for Four Attitude Items



E = Essay Test
OR = Oral Exam
TF = True-False Test
MC = Multiple-Choice Test
CMC = Computer Multiple-Choice
X = Frequency of *I* Scales

For the Time Pressure *J*-scale, there were again two dominant *I*-scales. One set of examinees found the computer multiple-choice test to have the least time pressure. In increasing order of time pressure, this was followed by TF, MC, OR, and E. The second group found TF to have the least time pressure, then MC, CMC, OR, and E.

The Anxiety *J*-scale showed the most variability in terms of test preference. No clearly dominant *I*-scales were evident, although there was a tendency for MC and CMC tests to have the lowest levels of anxiety for most of the examinees. On the Overall Preference *J*-scale, most of the examinees liked the MC tests best, the CMC tests second best, and the OR exams least.

Since score values were assigned to the examinees according to their positions on the four *J*-scales, a series of four separate multivariate analyses of variance were conducted between the *J*-scale score values and (1) low vs. high ability levels, (2) low vs. high number of items administered on the tailored test, (3) low vs. high amount of time spent taking the test, and (4) low vs. high levels of prior computer experience.

Overall significant findings resulted for only two of the analyses—number of items and amount of time spent taking the test, approximate F (4,51) = 2.58, $p < .05$ and approximate F (4,51) = 2.79, $p < .05$, respectively. However, none of the subsequent one-way analyses were significant in either case, making the findings difficult to interpret. It is clear that the attitudes of examinees taking many vs. few items or spending much vs. little time on the test differed significantly in terms of their scores on the four J-scale continua; however, where or how they differed is not clear. Perhaps part of the problem is related to the fact that none of the four J-scales was purely unidimensional. It was possible to fit only about half of the complete set of reported I-scales to any of the final four J-scales.

Preferences for CRT Displays

The final multivariate analysis of variance was performed on the results from the six Likert items dealing with the examinees' preference for the black-on-white compared to the white-on-black CRT display screens used for the tailored tests. The results presented in Table 3 indicate that the examinees' attitudes toward the two display formats were significantly different, approximate F (6,105) = 2.628, $p < .05$. The subsequent one-way analyses revealed that the differences occurred primarily on Items 1 and 5, F (1,110) = 5.48, $p < .05$ and F (1,110) = 8.34, $p < .01$, respectively. Both of these items refer specifically to reading difficulty experienced in taking the test. The examinees reported that the white lettering against a black screen background was significantly more uncomfortable on their eyes and was more difficult to read than the black-on-white screen format.

Table 3
Means and MANOVA for Attitudes

	Cell Means	
Attitude Scale Item	Black	White
1	3.125	3.679
2	3.750	3.821
3	3.382	3.436
4	3.945	3.982
5	3.564	4.036
6	3.055	3.109

Univariate Analyses of Variance

Source	*SS*	*df*	*MS*	*F*
Item 1--"Hurt Eyes"				
Black/White	8.58	1	8.58	5.48*
Error	172.34	110	1.57	
Item 5--"Reading Difficulty"				
Black/White	8.04	1	8.04	8.34**
Error	105.93	110	0.96	

**$p<.01$; $p<.05$

Discussion

One important, but not surprising, finding of this study was that the attitudes of examinees toward tailored tests were different in motivated test situations as compared to unmotivated test settings. If tailored achievement tests are to be used to classify or place individuals, to assign performance grades, or when a clear incentive for performance is present, the heightened anxiety levels of the examinees will be a factor. Of course, learning research suggests that heightened anxiety levels facilitate problem-solving performance for simple tasks, but inhibit performance on more complex tasks (Travers, 1972). Further research needs to be conducted in order to determine these effects in tailored achievement tests.

In this regard, it is interesting to note that the Coombs unfolding items yielded identical *J*-scales for the anxiety and overall test preference dimensions. This result may indicate that differences in anxiety levels for various types of tests dictated the examinees' overall preference levels for these tests. Thus anxiety may outweigh the effects of numerous other factors, such as time pressure or difficulty level, in terms of test preference.

In general, the tailored tests fared reasonably well compared to the other four test types measured on the *J*-scale attitude dimensions. In most cases the tailored tests were considered to be most similar to the traditional multiple-choice test format, which was the most preferred test type overall.

Finally, the CRT terminal with the white-on-black display screen was judged by the examinees to be significantly more difficult to read and harder on their eyes than the black-on-white display screen. Further research needs to be conducted in different settings and on different tailored testing tasks before it is possible to say whether or not reading difficulty actually interferes with test performance.

References

Betz, N. E., & Weiss, D. J. Psychological effects of immediate knowledge of results and adaptive ability testing (Research Report 76-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1976. (NTIS No. AD A027170)

Coombs, C. H. A theory of data. New York: John Wiley & Sons, Inc., 1964.

English, R. A., Reckase, M. D., & Patience, W. M. Application of tailored testing to achievement measurement. Behavior Research Methods and Instrumentation, 1977, 9, 158-161.

Hedl, J. J., O'Neil, H. F., & Hansen, D. N. Affective reactions toward computer-based intelligence testing. Journal of Consulting and Clinical Psychology, 1973, 40, 217-222.

Lushene, R. E., O'Neil, H. F., & Dunn, T. Equivalent validity of a completely computerized MMPI. Journal of Personality Assessment, 1974, 38, 353-361.

Reckase, M. D. Computerized achievement testing using the simple logistic model. Paper presented at the 1977 Annual Meeting of the American Educational Research Association, New York, NY, 1977.

Resmovic, V. The effects of computerized experimentation on response variance. Behavior Research Methods and Instrumentation, 1977, 9, 144-147.

Travers, R. M. W. Essentials of learning. New York: The Macmillan Company, 1972.

Weiss, D. J. Adaptive testing research at Minnesota--overview, recent results and future directions. Paper presented at the Conference on Computerized Adaptive Testing, Washington, DC, June, 1975.

REDUCTION OF TEST BIAS BY ADAPTIVE TESTING

STEVEN M. PINE
UNIVERSITY OF MINNESOTA

Because it has the capability of adapting to the specific characteristics of individual testees, computerized adaptive or tailored testing would appear to have potential for reducing test bias due to individual or group difference variables. These variables might include group differences in ability, motivation, test-taking anxiety, or tendency to either guess or omit items.

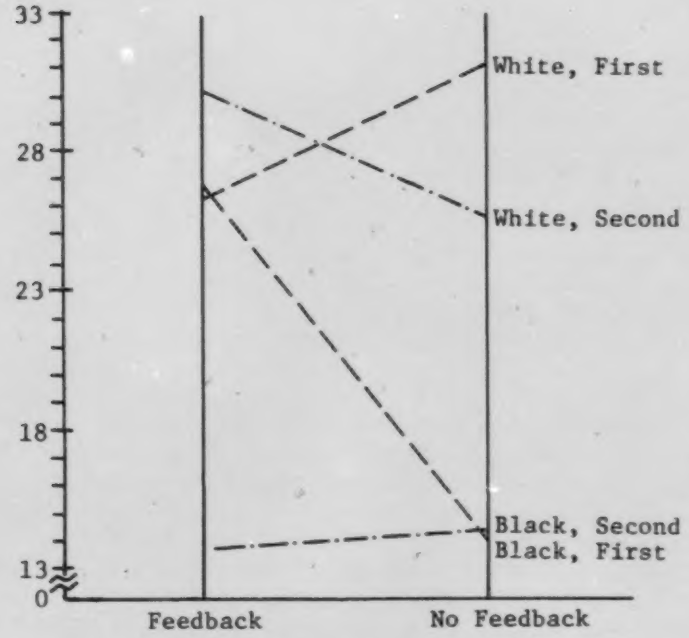
Previous research has provided some evidence for potential psychometric and psychological benefits to minority testees using computerized adaptive testing. Pine and Weiss (in press) demonstrated through a computer simulation that adaptive testing is able to reduce test unfairness and racial differences in test reliability. In a study conducted at the University of Rochester, Johnson and Mihal (1973) administered identical tests to black and white students in a conventional paper-and-pencil format and by computer; however, in this study the computerized test was not adaptive. White students scored significantly higher than blacks on the paper-and-pencil tests, but not on the computer-administered test.

In a second study, conducted at the University of Minnesota, Weiss (1975, p. 24) administered two computer-administered tests to about 100 high school students. The group was racially mixed, consisting of both white and black students. Both a conventional test and a pyramidal adaptive test were administered to each student; half of the group received the conventional test first, and half received the adaptive test first. In addition, half of the group received feedback after each item indicating whether or not their answers were correct; the other half received no feedback after each test item.

The data were analyzed for the conventional test only. Thus, the dependent variable in this analysis was number correct on the conventional test. The design was a 2×2×2 analysis of variance. The independent variables were (1) race--black and white; (2) feedback--immediate or none; (3) order--conventional test administered first or second in the pair. The results for the three-way analysis of variance showed that the only significant main effect was for race; however, there was a significant three-way order × race × feedback interaction.

As shown in Figure 1, when a conventional test was administered first under conditions of immediate feedback, the mean of the black students (26.38) was not significantly different from the mean of the white students (26.0)

Figure 1
Mean Scores for Blacks and Whites Completing
the 40-Item Conventional Test First and
Second, by Feedback Condition



completing the test under the same set of conditions. If this result can be replicated, it implies that race differences observed in test scores may be a function, not of differences in ability, but of differences in psychological effects of the conditions of administration. These findings, although not completely replicating those of Johnson and Mihal (1973), do support their general conclusion that conditions of test administration might affect motivational conditions, which in turn reduce race group differences to nonsignificant levels.

The purpose of the present study was to replicate and extend the previous findings that computerized testing can increase the test scores and the test-taking motivation of minority testees. Specifically, the present study compared a computerized adaptive test designed specifically to minimize test bias with a comparable paper-and-pencil test in order to determine possible racial differences on the following variables:

1. Ability test scores
2. Standard errors of measurement
3. Number of omitted responses
4. Test anxiety, motivation, and tendency to guess.

METHOD

Phase I

The study consisted of two distinct phases. Phase I was intended to develop and calibrate an item pool suitable for testing a group of racially mixed high school students. Three hundred and fifty high school students from two Minneapolis high schools with relatively large minority enrollments were tested. Using these data and data from a previous study, item characteristic curve parameters and an index of bias were calibrated for 250 vocabulary items.

Item Parameters

For each item the Phase I calibration produced estimates for two standard item characteristic curve (ICC) parameters and a third parameter, which was used to index bias. The two ICC parameters estimated were the discriminating power, a, and the item difficulty, b (the guessing parameter, c, was assumed to be equal to .2 for all items). Bias was indexed by an ICC version of Angoff and Ford's (1971) elliptical distance measure of item bias. This index is directly proportional to the difference between the item difficulties of two contrasted subgroups. In the present study this was the difference between the b values for the white and black subgroups.

Table 1
Examples of Vocabulary Item Types

Item Type	Example		
STANDARD	RESCUE	ILLEGAL	FEDERATION
	1. REMEMBER	1. FORBIDDEN	1. RESPECT
	2. REDUCE	2. DISTRESSING	2. ORGANIZATION
	3. MISTAKE	3. ENORMOUS	3. REPORT
	4. SAVE	4. LOYAL	4. GUARANTEE
	5. CHARGE	5. CHEAP	5. INFLATION
BLACK	DOZENS	RANKING	PULPIT
	1. BAKERS	1. MURDERING	1. PREACHING PROFESSION
	2. PERMITS	2. EXCHANGE OF INSULTS	2. ATTRACTIVENESS
	3. FALLS ASLEEP	3. PIG'S INTESTINES	3. QUARRY
	4. INSULTS	4. FRIED COW'S TAIL	4. PLANT TISSUE
	5. DONUTS	5. OLYMPIC EVENT	5. REDUCE
WHITE	BORSCH	TORTE	CAMEO
	1. OVERCOAT	1. CAKE	1. FLOWER
	2. DOG	2. TWIST	2. FRUIT
	3. PORTER	3. SHIRT	3. CRISIS
	4. SOUP	4. CRIME	4. CARVED FIGURE
	5. CHAMBER	5. ANSWER	5. DIAS

Item Types

The item pool consisted of five-alternative multiple-choice vocabulary items of distinct types. About one-half of the items were standard vocabulary items taken from the University of Minnesota and Educational Testing Service item files. The remaining items, however, were written especially for this study. Of these, thirty were purposely written to be biased against blacks. The remaining items were written specifically for blacks. These items were drawn from three main sources--black literature, two black lexicons, and the items written for this study by a black psychologist. Examples of each of the three item varieties are given in Table 1.

Phase II

In Phase II the calibrated item pool was used to form two standard paper-and-pencil tests and two computerized adaptive tests (CAT) administered on Cathode Ray Tubes (CRT's). Students had ample time to complete the tests and were instructed to guess if they could eliminate at least one alternative as incorrect.

Examinees

Two hundred and thirty students (half white and half black) were tested in Phase II. Of these, the data from 108 blacks and 107 whites were analyzed. Each student was given a McDonald's gift certificate worth $0.50 for participating in the study.

Figure 2
Assignment of Students to Experimental Conditions
for Phase II of Project MINISTEP

Group and Order	FB				NFB			
	BR		NBR		BR		NBR	
	P&P	CAT	P&P	CAT	P&P	CAT	P&P	CAT
Blacks								
Order 1	S_1 ⋮ S_{14}		S_{15} ⋮ S_{28}		S_{29} ⋮ S_{42}		S_{43} ⋮ S_{56}	
Order 2	S_{57} ⋮ S_{70}		S_{71} ⋮ S_{84}		S_{85} ⋮ S_{98}		S_{99} ⋮ S_{112}	
Whites								
Order 1	S_{113} ⋮ S_{126}		S_{127} ⋮ S_{140}		S_{141} ⋮ S_{154}		S_{155} ⋮ S_{168}	
Order 2	S_{169} ⋮ S_{182}		S_{183} ⋮ S_{196}		S_{197} ⋮ S_{210}		S_{211} ⋮ S_{224}	

Design

One of the paper-and-pencil tests and one of the computerized tests were designed to minimize test bias. This was referred to as the bias-reduced (BR) test. In addition, half of the students taking each test were given feedback (FB) after each item was administered, indicating whether or not their answer was correct. Each student took one computerized and one paper-and-pencil test. Half of the students took the computerized test first (Order 1), and half took the paper-and-pencil test first (Order 2). In addition, a test reaction questionnaire consisting of motivation, nervousness, guessing, and feedback scales was given to all students at the end of each test condition. Additional dependent measures in this study included three test performance measures indicating test score, standard error of measurement, and number of omitted items. The design of Phase II is shown in Figure 2.

Table 2
Item Characteristics of Items in Stradaptive Test Pool

Stratum	No. Items	Parameter	$\bar{X}$	*S.D.*	Minimum	Maximum
1	11	*a*	.86	.35	.410	1.487
		b	-1.78	.17	-2.039	1.571
		Bias	.74	1.11	-.247	3.730
2	19	*a*	1.00	.34	.531	1.775
		b	-1.12	.22	-1.488	-.909
		Bias	.34	.70	-1.197	.958
3	29	*a*	1.01	.33	.282	1.791
		b	-.58	.18	.889	-.306
		Bias	.53	.59	-.643	1.304
5	43	*a*	1.24	.45	.516	2.268
		b	.02	.18	-.278	.293
		Bias	.37	.51	-.675	1.263
5	35	*a*	1.03	.48	.087	2.215
		b	.59	.20	.315	.884
		Bias	.30	1.30	-5.463	3.526
6	17	*a*	1.13	.43	.447	2.080
		b	1.16	.15	.912	1.460
		Bias	.76	.63	-.519	1.570
7	17	*a*	.78	.70	.900	2.870
		b	3.28	3.60	1.520	16.838
		Bias	.54	3.28	-4.441	10.730

Test Instruments

Computerized adaptive tests. Both of the CAT tests studied--bias-reduced (BR) and non-bias-reduced (NBR)--used the stradaptive strategy developed by Weiss (1973). All items were sorted by difficulty level into one of seven strata. The two test conditions--BR and NBR--differed only in the way in which items were arranged within each stratum. For the non-biased-reduced (NBR) test conditions, items were arranged by item discrimination level, with the most discriminating items first. For the biased-reduced (BR) test condition, items were arranged by degree of item bias, with the least biased items first. The item characteristics of the stradaptive tests are summarized in Table 2.

An initial stratum assignment was made by asking each testee to rate him-self/herself on verbal ability on a three-point scale. The testee was then given the first item in the next-to-easiest, average, or next-to-most difficult stratum depending on his/her self-rating. If the testee's response to this first item was correct, he/she was branched to (administered an item from) the next more difficult stratum. If his/her response was incorrect, he/she was branched to the next easier stratum. If there was not a sufficiently easy or difficult stratum for the required branching (i.e., when an incorrect response was given to an item in Stratum 1, the least difficult stratum, or a correct response given to an item in Stratum 7, the most difficult stratum), the testee was given another item in the same stratum. In all cases the item administered was either the least biased item or the most discriminating (for the NBR condition) item remaining in the stratum.

Paper-and-pencil tests. The bias-reduced (BR) and non bias-reduced (NBR) paper-and-pencil tests were formed from 40 items not used in the stradaptive test item pool. These items were selected--20 from each of the stradaptive test pools--to have approximately the same average item discrimination and item bias as the first 10 items in the BR and NBR stradaptive test pools, respectively. It was impossible to exactly match item characteristics of the 20 items in the paper-and-pencil tests to the 20 used in the stradaptive test, since it could not be determined in advance exactly which 20 items would be administered in the stradaptive tests for each testee. The item characteristics of the paper-and-pencil tests are summarized in Table 3.

Table 3
Item Characteristics of Conventional Paper-and-Pencil Tests

	Bias Reduced			Non-Bias Reduced		
Statistic	*a*	*b*	Bias	*a*	*b*	Bias
Mean	1.57	.05	.80	1.02	.07	-.05
S.D.	.28	.71	.36	.47	.55	1.34
Minimum	1.166	-1.482	.219	.087	-1.482	-5.462
Maximum	2.268	1.136	1.709	2.268	1.009	.744

A special latent ink process was used to give feedback in the paper-and-pencil mode of administration. Students marked their answer sheets with a special pen which caused a latent image (previously invisible) to appear. The letter *Y* appeared if the correct answer was marked; the letter *N* appeared for incorrect answers.

Table 4
Test Reaction Questions for Nervousness and Motivation Scales

Scaled Score		Nervousness Scale
		WERE YOU NERVOUS WHILE TAKING THE TEST?
1	☐	NOT AT ALL
2	☐	SOMEWHAT
3	☐	MODERATELY SO
4	☐	VERY MUCH SO
		DID NERVOUSNESS WHILE TAKING THE TEST PREVENT YOU FROM DOING YOUR BEST?
4	☐	YES, DEFINITELY
3	☐	YES, SOMEWHAT
2	☐	PROBABLY NOT
1	☐	DEFINITELY NOT
		Motivation Scale
		DID YOU CARE HOW WELL YOU DID ON THE TEST?
4	☐	I CARED A LOT
3.2	☐	I CARED SOME
2.4	☐	I CARED A LITTLE
1.6	☐	I CARED VERY LITTLE
.8	☐	I DIDN'T CARE AT ALL
		DID YOU FEEL CHALLENGED TO DO AS WELL AS YOU COULD ON THE TEST?
1	☐	NOT AT ALL
2	☐	SOMEWHAT
3	☐	FAIRLY MUCH SO
4	☐	VERY MUCH SO
		WERE YOU INTERESTED IN KNOWING WHETHER YOUR ANSWERS WERE RIGHT OR WRONG?
4	☐	I WAS VERY INTERESTED
3	☐	I WAS MODERATELY INTERESTED
2	☐	I WAS SOMEWHAT INTERESTED
1	☐	I DIDN'T CARE AT ALL

Table 5
Test Reactions Questions for Guessing and Feedback Scales

Scaled Score		Guessing Scale
		ON HOW MANY OF THE QUESTIONS DID YOU GUESS?
4	☐	ALMOST ALL OF THE QUESTIONS
3.33	☐	MORE THAN HALF OF THE QUESTIONS
2.67	☐	ABOUT HALF OF THE QUESTIONS
2	☐	LESS THAN HALF OF THE QUESTIONS
1.33	☐	ALMOST NONE OF THE QUESTIONS
.67	☐	NONE OF THE QUESTIONS
		HOW OFTEN WERE YOU SURE THAT YOUR ANSWERS TO THE QUESTIONS WERE CORRECT?
.8	☐	ALMOST ALWAYS
1.6	☐	MORE THAN HALF OF THE TIME
2.4	☐	ABOUT HALF OF THE TIME
3.2	☐	LESS THAN HALF OF THE TIME
4	☐	ALMOST NEVER
		Feedback Scale
		DID RECEIVING FEEDBACK AFTER EACH QUESTION INTERFERE WITH YOUR ABILITY TO CONCENTRATE ON THE TEST?
1	☐	NO, NOT AT ALL
2	☐	YES, SOMEWHAT
3	☐	YES, MODERATELY SO
4	☐	YES, VERY MUCH SO
		DID GETTING FEEDBACK AFTER EACH QUESTION MAKE YOU NERVOUS?
1	☐	NO, NOT AT ALL
2	☐	YES, SOMEWHAT
3	☐	YES, MODERATELY SO
4	☐	YES, VERY MUCH SO

Test reaction questions. The psychological reactions to each testing condition were assessed by administering test reaction questions consisting of four factor-analytically-derived scales designed to measure (1) nervousness, (2) motivation, (3) tendency to guess, and (4) reaction to feedback. (See Prestwood & Weiss, 1977, for a description of the derivation of these scales.) Tables 4 and 5 show the test reaction items by scale. Items were administered to each testee twice--once after each testing condition. Testees in the no feedback condition were given only the motivation, nervousness, and guessing scales.

Test Performance Measures

Three test performance measures, indicating ability test score, standard error of measurement, and number of omitted responses were investigated. The ability test score was obtained by a Bayesian scoring procedure similar to the one developed by Owen (1975). This procedure provided a means of generating comparable scores for both the conventional and adaptive tests. The posterior Bayesian variance was used as an estimate of the standard error of measurement; Jensema (1974) has indicated the relationship between these measures.

RESULTS AND DISCUSSION

Test Performance Measures

Significant *F* ratios for the 2×2×2×2×2 repeated measures analysis of variance performed on the Bayesian ability scores, the Bayesian variance, and the number of omitted responses are given in Table 6.

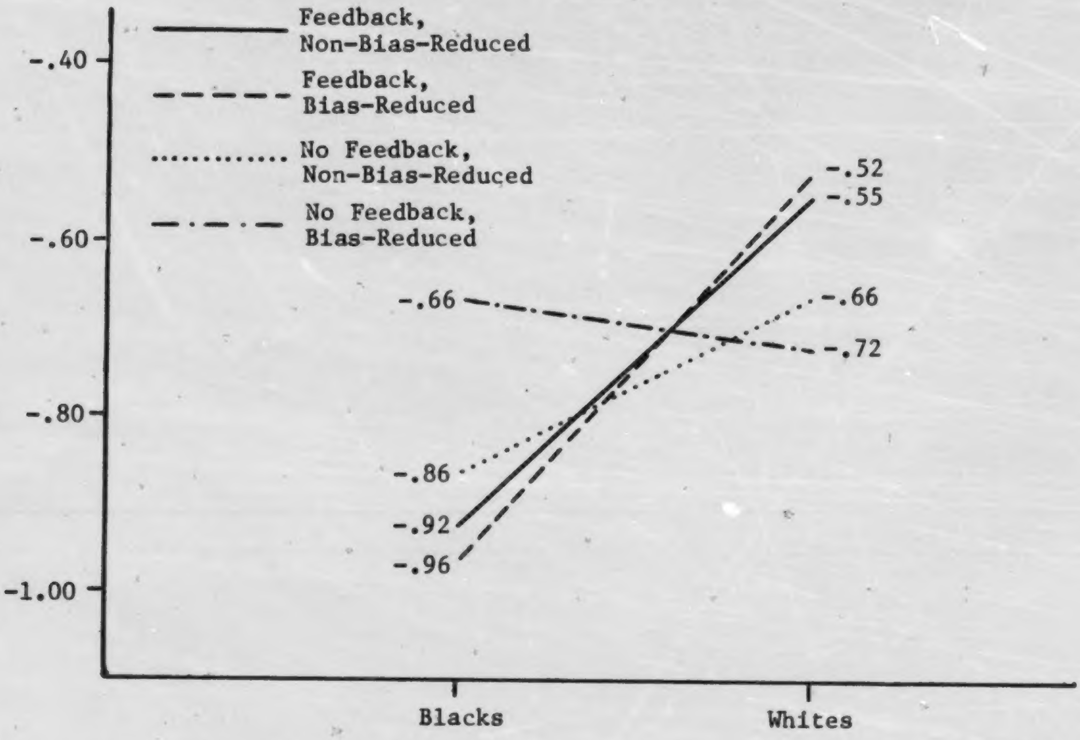
Table 6
Significant *F* Ratios for ANOVAs on Performance Measures

Dependent Variable	Source	Degrees of Freedom	Mean Square	*F*	*p*
Bayesian Score	R	1	5.83	6.60	.011
	RFB	1	2.93	3.31	.070
	Error (R,RFB)	199	.88		
	MRFB	1	.56	3.65	.057
	Error (MRFB)	199	.53		
Bayesian Variance	B	1	.46	582.28	.000
	ROB	1	.00	4.69	.031
	Error (B,ROB)	199	.01		
	MO	1	.02	13.17	.000
	MB	1	.01	6.01	.015
	MOB	1	.01	9.32	.003
	Error (MO,MB,MOB)	199	.01		
Number of Omits	F	1	96.27	4.43	.037
	RF	1	90.10	4.15	.043
	Error (F,RF)	199	21.72		

Bayesian ability estimates. The only significant ($p<.02$) main effect on the Bayesian ability scores was for race (R), with whites scoring significantly higher than blacks. This is the result commonly found in the testing literature. It is, however, particularly noteworthy in the present study, since one of the test conditions (BR) was specifically designed to reduce test score differences and included "Black-type" items. This result cannot alone be interpreted as indicating that the BR condition was ineffective, since the size of this effect may have been larger had the BR condition not been included.

Racial effects resulting from the manipulation of the bias-reduction (B), feedback (F), and mode (M) variables can be interpreted in the significant interactions involving the race variable. Two such interactions were significant. The race × feedback × bias-reduced interaction, which is shown in Figure 3, is particularly interesting.

Figure 3
Race × Bias-Reduced × Feedback Interaction for Bayesian Scores



The influence of feedback was straightforward for whites. Whites performed better on both tests (BR and NBR) when feedback was provided. The results were different for blacks. Blacks performed worse on both tests when feedback was given; but when feedback was not given, they scored better on the bias-reduced test.

These results may reflect race differences in the rate and aversiveness of negative feedback. Blacks received more negative feedback than did whites because they tended to answer more items incorrectly. Furthermore, the negative feedback which blacks received in the bias-reduced condition was likely to be particularly aversive, since they were being told they did not know the meaning of "black-type" words. The four-way race × bias-reduced × feedback × mode interaction indicated that although blacks scored lower when feedback was given, they did relatively better when the items were administered by computer with the adaptive testing strategy.

Bayesian posterior variance. The analysis of variance performed on the Bayesian posterior variance dependent measure produced a significant main effect for the type of test (BR or NBR) and for the race × order × bias-reduced and mode × order × bias-reduced interactions. It should be recalled that (1) the Bayesian posterior variance provides an estimate of the standard error of measurement and (2) the only distinction between the BR and NBR stradaptive tests was the order in which the items were arranged within each stratum. In the BR tests, items were arranged from least biased to most biased; in the NBR tests, items were in descending order of their item discriminations. Therefore, these results are consistent with the fact that the standard error of measurement at a fixed ability level is decreased by increasing item discrimination. The mode × bias-reduced interaction provided additional support to the growing body of research (Vale & Weiss, 1975a, 1975b; Betz & Weiss, 1976a, 1976b), showing that stradaptive tests produce smaller standard errors of measurement than comparable conventional tests.

Number of omits. The administration of feedback significantly ($p<.05$) reduced the number of omitted items. The significant two-way interaction between race and feedback shown in Figure 4 indicates the nature of the feedback effect.

Figure 4
Mean Number of Omits as a Function
of Feedback for Blacks and Whites

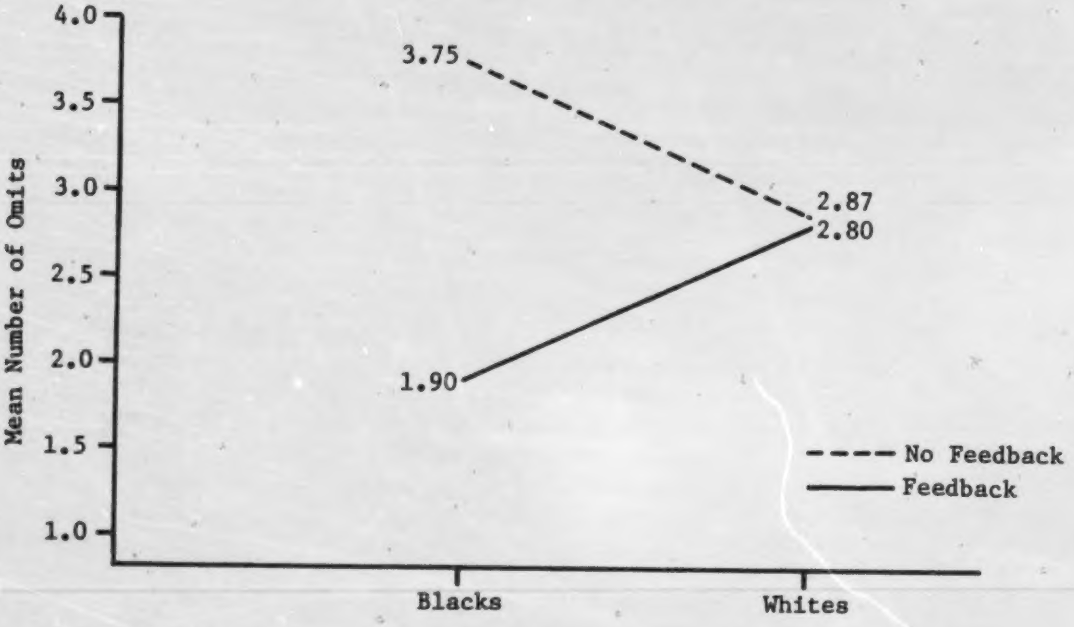


Figure 4 indicates that the significance of the feedback effect was almost entirely attributable to its influence on blacks. When feedback was administered, blacks omitted significantly ($p<.05$) fewer items. However, this reduction in the tendency to omit items did not lead to a significant increase in test scores for blacks. This is probably because a decrease in the number of omits was accompanied by an increase in guessing, which would not be expected to significantly increase a Bayesian ability estimate.

Test Reaction Scales

Table 7 gives the results of the analysis of variance for the test reaction scales. Significant *F* ratios were obtained for the feedback, nervousness, motivation, and tendency-to-guess scales.

Table 7
Significant *F* Ratios for ANOVAs on Test Reaction Scales

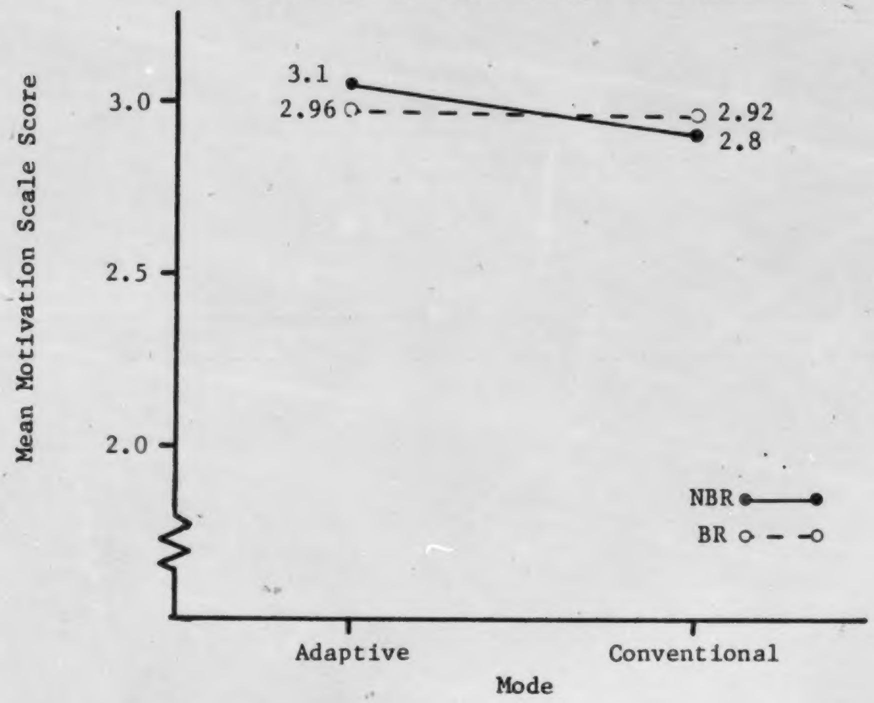
Dependent Variable	Source	Degrees of Freedom	Mean Square	*F*	*p*
Feedback	R	1	7.78	8.60	.00
	Error (R)	80	.91		
	MO	1	.87	4.13	.05
	Error (MO)	80	.21		
Nervousness	OB	1	5.37	7.46	.01
	ROFB	1	3.17	4.41	.04
	Error (OB,ROFB)	185	.72		
	M	1	1.22	5.01	.03
	MB	1	.87	3.57	.06
	Error (M,MB)	185	.24		
Motivation	RO	1	4.68	5.41	.02
	ROFB	1	5.10	5.89	.02
	Error (RO,ROFB)	185	.87		
	M	1	2.17	14.04	.00
	MB	1	1.25	8.09	.01
	MROB	1	.83	5.35	.02
	Error (M,MB,MROB)	185	.15		
Guessing	OFB	1	2.74	3.67	.06
	Error (OFB)	185	.75		
	M	1	2.06	6.06	.02
	MRO	1	1.26	3.72	.06
	MOB	1	1.53	4.51	.04
	Error (M,MRO,MOB)	185	.34		

Feedback scale. The significant ($p<.005$) race (R) effect for the feedback scale indicated that blacks were more adverse to receiving feedback than were whites. They felt that receiving feedback impaired their ability to concentrate and made them somewhat nervous. This finding is consistent with the negative feedback hypothesis and may, at least partially, explain why blacks scored lower than whites.

Nervousness scale. The only significant ($p<.05$) main effect for the nervousness scale indicated that testees, in general, were more nervous when tested by an adaptive testing strategy than when tested conventionally (M effect). The only significant effect ($p<.05$) involving race was the four-way race × order × feedback × bias-reduced interaction (ROFB). Because of the number of terms involved, it is difficult to interpret the exact meaning of this interaction. One explanation of the main effect, however, can be attributed to the tendency of many people to become nervous when required to work in a novel environment, particularly when that environment involves the operation of a machine. Another explanation can be attributed to the nature of the adaptive testing procedure. In adaptive testing, item difficulties are tailored to the ability level of each testee. Therefore, testees are constantly being challenged to the limit of their ability. This would understandably increase nervousness or test anxiety.

Motivation scale. Again, the only significant ($p<.001$) main effect was due to mode of test administration (M), indicating higher test-taking motivation in the adaptive testing condition. The significant ($p<.005$) mode × bias-reduced (MB) interaction (see Figure 5) shows that motivation did not vary as a function of the biasedness of the test in the adaptive testing condition, but did in the paper-and-pencil condition. These results support results in the adaptive testing literature (Weiss & Betz, 1973; Weiss, 1974) that test-taking motivation can be increased with adaptive testing.

Figure 5
Mean Motivation Scale Scores as a Function of
Mode of Administration and Bias-Reduction



Significant race effects emerged in the race × order (RO), race × order × feedback × bias-reduced (ROFB), and mode × race × order × bias-reduced (MROB) interactions. In conjunction with the fact that the race × order × feedback × bias-reduced interaction was also found to be significant on the nervousness scale, there appears to be sufficient evidence to indicate a racial difference in the psychological aspects of test-taking behavior.

Guessing scale. Once again, the only significant ($p<.02$) main effect was for mode of administration (M), with a lesser tendency to guess in the adaptive testing condition. This finding is consistent with the fact that adaptive tests tend to administer fewer items above a testee's ability level; consequently, there is less reason for an individual to guess.

Summary and Conclusions

The present study provided general information on the relative merits of computerized adaptive versus conventional paper-and-pencil testing and provided specific information on the use of these two methods for testing minority examinees. The results of the study add to the growing body of research which has shown the general superiority of adaptive testing over conventional paper-and-pencil testing. It was found that in comparison to paper-and-pencil tests, computerized stradaptive tests (1) produced ability estimates with smaller standard errors of measurement, (2) reduced the tendency to guess, and (3) increased test-taking motivation.

Evidence also emerged which indicated that people are anxious (nervous) following an adaptive test. This finding is understandable, since in adaptive testing,items are tailored to the ability level of each testee and will therefore tend to challenge the testee to the limits of his/her ability. This increased nervousness did not, however, decrease performance or reduce the accuracy of test scores.

No definitive evidence was manifested to support the claims that either feedback and/or adaptive testing significantly improves the test performance of blacks; nor did the bias-reducing condition produce a significant increase in black test scores. Several important factors about the conditions of this study, however, must be kept in mind when interpreting these results. First, of the 20 items administered to each student,only about 5 were likely to be words more common in black culture. In view of the relatively small cell frequencies (between 12 and 15), it is unlikely that there was sufficient power to detect a test score increase for blacks attributable to approximately five items. Furthermore, blacks did score nonsignificantly higher overall on the bias-reduced tests and on the adaptive bias-reduced tests, thereby suggesting that the conditions provided by these tests may have been beneficial to minority testees.

Secondly, in order to evaluate the extent to which these results support the feedback effect reported by Weiss (1975, p. 35), several important differences between these studies must be considered. The Weiss study did not include a bias-reduced condition and gave feedback which was specifically chosen to be meaningful to blacks. (The feedback in the present study consisted of a simple "correct", or "incorrect.") In those conditions of the present study corresponding most closely to the conditions of the Weiss study, a similar feedback effect was found--blacks scored relatively higher ($p<.15$) when they were given feedback on the computer. Furthermore, test-taking motivation was significantly higher for blacks overall and was nonsignificantly higher for blacks in the computerized feedback condition. Also, blacks omitted significantly fewer items than did whites when feedback was given.

The apparently contradictory finding that blacks scored significantly lower when given feedback on the computer occurred only in the bias-reduced condition. Blacks may have been particularly distressed upon being told that they didn't know the meaning of words common to their culture. This hypothesis is supported by blacks reporting that getting feedback made them significantly more nervous and somewhat reduced their ability to concentrate.

The finding from the test reaction scale concurs with the general conclusions of the Johnson and Mihal (1973) and Weiss (1975) studies in demonstrating that blacks react differently to the conditions of testing (e.g., feedback and mode of test administration) than do whites. Although race test score differences were not reduced to nonsignificant levels in the present study, the direction of test score differences was consistent with the conclusions of previous studies. This suggests that racial difference in test scores can be reduced with computerized testing techniques.

References

Angoff, W. H., & Ford, S. F. Item-race interaction on a test of scholastic aptitude (Research Bulletin RE-71-59). Princeton, NJ: Educational Testing Service, 1971.

Betz, N. E., & Weiss, D. J. Effects of immediate knowledge of results and adaptive testing on ability test performance (Research Report 76-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, June 1976. (NTIS No. AD A027147) (a)

Betz, N. E., & Weiss, D. J. Psychological effects of immediate knowledge of results and adaptive ability testing (Research Report 76-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, June 1976. (NTIS No. AD A027170) (b)

Jensema, C. J. An application of latent trait mental test theory. British Journal of Mathematical and Statistical Psychology, 1974, 27, 20-48.

Johnson, D. I., & Mihal, W. M. Performance of blacks and whites in computerized versus manual testing environments. American Psychologist, 1973, 28, 694-699.

Owen, R. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of the American Statistical Association, 1975, 70, 351-356.

Pine, S. M., & Weiss, D. J. A comparison of the fairness of adaptive and conventional testing strategies (Research Report 78-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, in press.

Prestwood, J. S., & Weiss, D. J. Accuracy of perceived test-item difficulties (Research Report 77-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, May 1977 (NTIS No. AD A041084)

Vale, C. D., & Weiss, D. J. A study of computer-administered stradaptive ability testing (Research Report 75-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, October 1975. (NTIS No. AD A018758) (a)

Vale, C. D., & Weiss, D. J. A simulation study of stradaptive ability testing (Research Report 75-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, December 1975. (NTIS No. AD A020961) (b)

Weiss, D. J. The stratified adaptive computerized ability test (Research Report 73-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, September 1973. (NTIS No. AD 768376)

Weiss, D. J. Strategies of adaptive ability measurement (Research Report 74-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, December 1974. (NTIS No. AD A004270)

Weiss, D. J. Adaptive testing research at Minnesota: Overview, recent results, and further directions. In C. L. Clark (Ed.), Proceedings of the First Conference on Computerized Adaptive Testing. Washington, DC: U.S. Civil Service Commission, 1976.

Weiss, D. J., & Betz, N. E. Ability measurement: Conventional or adaptive? (Research Report 73-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, February 1973. (NTIS No. AD 757788)

Acknowledgments

This research was supported by contract N00014-76-C-0244, NR No. 150-383, with the Personnel and Training Research Programs, Office of Naval Research. The author is indebted to Frederick Lord of the Educational Testing Service for providing many of these items and also wishes to thank Harvey Linder and Ken Jones for writing many of the items used in this study.

DISCUSSION: SESSION 3

NANCY S. COLE
UNIVERSITY OF PITTSBURGH

The studies of Prestwood, Koch and Patience, and Pine were concerned with the conditions of the testing situation and the effects on people's performance. The area is an important one for both traditional standardized tests and for adaptive tests, and it has received too little attention in the past.

Prestwood's paper was concerned with the effect of varying proportion correct, test strategy, and knowledte of results on estimates of examinee ability. This study raised the following questions:

1. The motivation of subjects is a natural concern when the research subjects are introductory psychology students participating for course credit. The question arises whether or not there appeared to be a motivational problem. (Testee motivation is an important concern which was raised again in the Koch and Patience study.)
2. Prestwood never discussed a primary result of the study--that testees did significantly better on the stradaptive test. There seems to be no explanation for that result, since the stradaptive test in this case was not a typical one which adapted to the individual's ability level but one which adapted to the pre-determined difficulty level. When the primary main effect found seemed to make no sense, I was somewhat uncomfortable with the data. The study would probably have been improved by not varying the testing strategy, since there was no particular reason for the variation.
3. The two-way interaction (knowledge of results and difficulty), although interesting, does not seem to conform very well to the Betz and Weiss studies. The largest discrepancy in the ability estimates was for differing difficulty in the conditions in which there was no knowledge of results: Students did better (had higher ability estimates) with an easy test than with a difficult test. This agrees with the conventional test development wisdom of not making a test too difficult to get maximal performance.
4. The results of the attitude measures do not mesh especially well with the ability estimate results. For example, motivation was higher under knowledge of results, but not ability estimates.

The Koch and Patience study on student attitudes was derived from a conglomeration of studies which would, perhaps, be of more interest than their results.

1. The most important finding is probably the suggestion of different attitudes about time pressure and anxiety under motivated rather than

non-motivated conditions. I like their consideration of a test situation in which test scores are important to the student. However, the dependent variables are of limited interest since so little is known about how or if they translate into test performance effects. It would be preferable, for example, to replicate Prestwood's study under such motivativational conditions.

2. The attitudes toward different types of tests are potentially interesting if people clearly prefer different types. That preference might then be used to match people to testing conditions. However, the students most preferred the comfortable and familiar multiple-choice test and least preferred the oral exam (which always has a poor reputation among students). Thus, the agreement of the students in their preference for the familiar test situation reduced the interest of these results.
3. Adaptive achievement testing needs clearer direction than it had in this study. Were the students to have objected to a 10-item achievement test, it would be totally understandable, since achievement needed to be tested in a large number of homogeneous content domains--not 10 items across a diffuse domain.

Pine's paper on reducing test bias was frustrating. He went to the effort of writing potentially biased items (which few people have tried to do), and he tested them in relevant groups of high school students. Yet he erred at the item calibration stage (about which he did not tell us); and one, therefore, does not know what Pine found in the study. I should like to further address the following specific points:

1. The study concerning computer simulation as an effective way to demonstrate how bias could be reduced should be either explained in the text or dropped.
2. Phase I of the study--the calibration of the 250 items and the computation of bias indices--could be the most interesting part of the study. Pine expects to report it elsewhere, but it would seem that the Phase II study cannot be properly interpreted without a clear understanding of the item parameters from Phase I and the nature and characteristics of the items actually used on the bias-reduced test. There are several problems here, including the dimensionality of the mixed pool. One of the consistent findings of item bias studies is the finding that about 5% of the items in a pool are significantly biased; yet it is not possible to see biased content in the items so identified. The statistical bias indices seem to be greatly capitalizing on chance. A student of mine studied the same test both in several different groups and in the same group over subsequent years. A different set of items was labelled biased each time. Many of the items Pine used in the bias-reduced test seem to involve the same sort of chance variation in the pro-black direction. If so, the bias-reduced test was not as bias reduced as Pine thinks. On this same point, the extent to which the specially written pro-black items will contribute to the ability estimates is doubtful. If blacks have lower discriminations, that would account for some of Pine's results.

3. The result of a race effect is not surprising if the bias-reduced test was constructed mainly of items showing chance deviations in favor of blacks, which would not be repeated in this sample.
4. The estimation of ability in this paper contrasts with that used in Prestwood's paper; it would be interesting to compare the two approaches on both sets of data.
5. Because this study used a very complex design, it is very difficult to interpret results unless one knows quite clearly what was being looked for. There was no special theoretical direction to the study. It might have profited from eliminating the mode and/or feedback conditions. It would seem that these complex interactions are very difficult to explain; one first needs a better understanding of simpler effects under more restricted conditions.
6. Pine's result about blacks being negative to feedback and obtaining lower mean scores with feedback under some conditions contradicts the Weiss study; probably neither's results should be taken too seriously without further replication. It certainly would be interesting and important were some consistent results to emerge.
7. Finally, there were several possible definitions of bias. In the original item pool, Pine defined bias in terms of the difference of the difficulty parameters. It could have been defined in terms of the difference in both item parameters. In the results the conditions could be considered as biasers when they led to different ability estimates for people. In this sense, feedback could be viewed as a bias inducer in Pine's results. This same viewpoint could be applied to Prestwood's study. Conditions are potential biases in the sense that they can produce consistent underestimates of ability.

All of the studies would profit from additional thought concerning the direction and objectives of this line of research. Is the goal to find conditions which minimize differential effects on ability estimates? to find conditions which maximize ability estimates in general? or conditions which maximize ability estimates for each individual? Is there a need to consider characteristics of the individual to interact with testing conditions to produce maximal ability estimates? And, finally, do the conditions contribute valid or irrelevant variation?

SESSION 4

PERFORMANCE TESTING BY INTERACTIVE SIMULATION

THE USE OF SIMULATION IN PERFORMANCE TESTING

CHRISTINE MCGUIRE
UNIVERSITY OF ILLINOIS
MEDICAL CENTER

PROBLEMS OF PERFORMANCE MEASUREMENT IN COMPUTER-BASED SIMULATION

BRUCE W. KNERR
ARMY RESEARCH INSTITUTE

THE DECISION MEASUREMENT SYSTEM AS A MEANS OF TESTING PERFORMANCE BY SIMULATION

M. A. ROBINSON
DATA DESIGN LABS
C. L. WALKER
NAVAL GUIDED MISSILE SCHOOL

DISCUSSION

ERNST Z. ROTHKOPF
BELL LABORATORIES

SESSION 4: ABSTRACTS

THE USE OF SIMULATION IN PERFORMANCE TESTING

CHRISTINE McGUIRE

In most occupations and professions, simulation techniques will yield more valid, reliable, and representative performance data than can be obtained from either real life observations or more conventional methods of testing. Written, computer-administered, and oral-interactive simulations may be used to test data-gathering and decision-making skills; various media (e.g., visual representations, sound simulators, and three-dimensional models) may be used to test observation and interpretative skills; oral-interactive simulations may be used to test interpersonal and communication skills; and three-dimensional models with or without computer capabilities may be used to test technical skills. These simulation devices are described and illustrated, and their advantages and disadvantages are discussed in relation to their present use in assessing health professionals.

PROBLEMS OF PERFORMANCE MEASUREMENT IN COMPUTER-BASED SIMULATION

BRUCE W. KNERR

The Adaptive Computerized Training System (ACTS) is a computer-based system which uses adaptive techniques to train electronic troubleshooting procedures. Using ACTS, the student troubleshoots an electronic circuit which is simulated by ACTS. An adaptive program, based on an expected utility decision model, "learns" the student's utility structure. The student utilities can then be compared with those of an expert, and instructional feedback and the problem presentation sequence can be based on the discrepancies between the two sets of utilities. Evaluations of ACTS indicate that the utility estimation algorithms operate correctly and that student performance on the ACTS improves with practice, even in the absence of any utility-based feedback. Full utilization of ACTS potential requires the resolution of several measurement issues regarding methods of adapting instruction on the basis of student utilities: (1) the degree of similarity among expert utility structures, (2) correspondence between utility structures and troubleshooting sequences, and (3) possibilities of reducing the size of the utility matrix.

THE DECISION MEASUREMENT SYSTEM AS A MEANS OF TESTING PERFORMANCE BY SIMULATION

M. A. ROBINSON AND C. L. WALKER

The Decision Measurement System (DMS) is a multiple-choice skill-related test that uses latent image techniques to control branching between items. It provides a realistic simulation of electronic equipment by presenting panel displays, waveforms, and other diagnostic information in a Data Book that is referenced by the test items. The DMS is being used by the Navy to assess the ability of submarine technicians to follow established maintenance procedures, as well as to identify and locate equipment faults for which the troubleshooting procedures are not well documented. This paper describes the circumstances that led the Navy to adopt this form of testing, the method for scoring the DMS, and the evidence for the validity of this test. In addition, research in the application of a task taxonomy to the items of the DMS is discussed.

THE USE OF SIMULATION IN PERFORMANCE TESTING

CHRISTINE MCGUIRE
UNIVERSITY OF ILLINOIS AT THE MEDICAL CENTER

In the newer and more hazardous occupations, simulation has been the technique of choice for both instruction and assessment. Simulation has held this preferred position because it has been recognized that while practical experience and performance testing are essential, exclusive reliance on real situations for teaching and testing is either dangerous (e.g., for aviators) or unfeasible (e.g., for astronauts). In fields such as these, simulation presents an attractive alternative both to more conventional instructional and evaluation techniques and to "on-the-job" teaching or testing. It is attractive because it provides the capability for imitating critical aspects of reality, while at the same time freeing both teacher and student (or examiner and examinee) from the risks and inconveniences always present in the real world.

Reduced to its essence, testing which employs true simulation technique is characterized by four essential characteristics:

1. The test situation imitates one or more critical aspects of reality; it does not pretend to duplicate that reality in its entirety.
2. The test situation requires the active participation of the respondent.
3. Actions of the respondent may modify the situation.
4. Whether or not the situation is modified, the respondent's participation can trigger appropriate feedback which may be utilized for subsequent decisions about pending action.

Limitations and Advantages of Simulation

It should be obvious that although these characteristics make simulation an attractive approach to testing, they also limit the areas of its usefulness. First, there are some aspects of reality that cannot at the present time be economically simulated, if they can be simulated at all. Second, there are some components of competence which cannot, or should not, be measured in a simulated situation. For example, simple recall of factual information is more economically and directly measured by conventional techniques of objective testing. At the other extreme, professional habits can be assayed only by careful and repeated observation over a long period of time in diverse settings. Between these two extremes simulation provides a number of important advantages.

Perceived Relevance

Compared with more conventional evaluation strategies, simulation offers the possibility of designing test situations so that they correspond more closely to the life situations a professional actually faces. Tests and summative evaluations composed of such exercises appear far more relevant than conventional examinations. Such perceived relevance is, at the very least, psychologically beneficial; and in these days of concern about the cultural biases of our tests, perceived relevance may turn out to be our only legal defense.

Predetermination and Preselection of the Task

It is also of prime importance that by utilizing simulation techniques, perceived relevance can be achieved without being dependent on the accidents of nature and the flow of real problems extant at the particular place and the specific moment when an assessment is to be made. It is obvious that simulation makes it possible to predetermine precisely the exact task which examinees are to be required to perform. Furthermore, it is clear that by eliminating the "noise" always present in reality, simulation makes it possible to focus on the elements of primary concern in a testing situation. Thus, irrelevant and confusing complexities that would contaminate the assessment can be eliminated. In addition, the fact that assessment tasks can be preselected and predesigned means that they can be carefully graduated in difficulty to be maximally appropriate for a particular target population. Finally, by developing a standard set of parallel simulations, it is possible to confront examinees repeatedly with interesting variations of what is essentially the same task, until there is clear evidence that each has mastered it.

Standardization of the Task

Simulation enables an examining body to standardize the task for all examinees. In the case of medical examinations, through simulation all examinees can be given exactly the same problem without subjecting any patients to repeated harassment.

Improved Sampling of Performance

By standardizing the tasks and focusing on the most significant aspects in each, it is possible in a given time period to sample an individual's performance with respect to a much broader and more representative group of problems. This, of course, requires careful selection and definition of the tasks. But this capability of broad and representative sampling of performance can rarely be accomplished within a reasonable time frame in the observation of non-simulated performance.

Furthermore, with simulation it is possible to test a candidate's ability to respond to an emergency or to handle a slowly evolving situation, without waiting years for the "right" emergencies to occur or spending a lifetime watching the gradual deterioration in a real situation. In carefully developed simulations any problem (ranging from the most urgent emergency to the most glacial evolution over eons) can be collapsed into a half-hour exercise and then summoned on demand.

Increased Accuracy of Performance Ratings

When the exact tasks that are to compose an examination are precisely defined and preselected, it is possible to develop specific detailed criteria for judging examinee performance. It is then also possible to train examiners to apply these criteria equitably, consistently, and in such a manner that feedback is provided which is of maximum benefit both to the student and to those who must judge his or her readiness for promotion, graduation, licensure, or certification. Because this type of examiner training is feasible, there is no question that a higher degree of inter-rater reliability can be achieved in the scoring of examinee performance on simulated problems than on real problems.

Increased Responsibility and Realistic Feedback in a Practical Time Frame

One of the most important advantages of simulation over reality consists in the fact that examinees can be allowed full responsibility, incurring the possibility of disaster without any risk other than to a piece of paper, a computer or his or her own psyche. This is important not only in teaching, but also in testing. For example, subjected to a neophyte health professional, a simulated patient can (1) become progressively sicker, (2) develop iatrogenic complications, (3) commit suicide, or (4) die "naturally" and still be resurrected repeatedly, to tirelessly serve in the retesting of the same or other candidates. And at every stage of the evolving problem, the examinee can be provided with instructive feedback about the patient's response to each intervention.

Increased Learning

The nature of the feedback provided in simulations compared with that available in either conventional testing or in the real world should be noted. While the former rarely provides any genuinely relevant feedback, the latter often furnishes it only indirectly and equivocally, with long delay. In contrast, the prompt, specific, and unambiguous feedback characteristic of well-designed simulations makes them a powerful tool not only in assessment, but also in the enhancement of learning. This is an important consideration in the choice of any evaluation strategy.

Educational Impact

Finally, one of the main advantages of simulation is its educational impact. The feedback that occurs in a simulated situation is itself instructive and can thus contribute significantly to improved performance. Beyond this, however, it is important to note that tests are exceedingly potent instruments for directing the attention and efforts of thousands of students, trainees and--in the future, perhaps--practicing professionals toward that which is regarded as important by the individuals or boards who are authorized to certify competence in a profession. Effective simulations have the power to direct that attention toward developing those aspects of competence which are valued by a profession.

Types of Simulations

For these reasons simulation technique is being employed increasingly in the licensure and certification of physicians and other health professionals. It has been found to be of special value in the measurement of four aspects of professional competence: (1) observational and interpretive skills, (2) clinical judgment, (3) interpersonal skills, and (4) technical skills.

Observational and Interpretive Skills

In the assessment of observational and interpretive skills, it is clear that provision must be made for the testing of the physician's ability to note and to interpret data that become available through many sensory modalities (e.g., visual data, auditory data, or data from palpation and percussion). Consequently, to assess these important skills, simulated clinical and laboratory data are presented in realistic form; and the examinee is simply asked to interpret them.

These simulations may be based on data, such as photographic reproductions of X-rays, gross and microscopic specimens, lesions, ECG or EEG tracings; these are presented in a form that imitates life as closely as possible. In addition, simulations may be based on sound simulators which have been developed for playback of high fidelity recordings of heart, breath, and abdomial sounds through individual stethophones. Data involving a combination of sound, color, and movement are presented in high quality movies or videotapes depicting the physician-patient interview and/or certain aspects of the physical examination. More recently, several three-dimensional models have been developed into which varied pathology can be inserted. With these models the aspiring physician can be required to demonstrate both skill in eliciting the findings and ability to interpret them. Using any or all of these modalities, simulation exercises can be constructed that sample examinees' abilities to (1) describe precisely the findings presented in auditory, visual, or other forms, (2) interpret the significance and implications of these findings, (3) anticipate other findings that might be related to those demonstrated, and (4) predict what might exacerbate or relieve the symptoms.

When oral or short answer formats are employed for examinations based on simulated sights and sounds, it is possible, (though laborious) not only to code each response as correct or incorrect, but also to identify the kinds of errors a candidate has made. Thus, detailed profile scores can be reported for each candidate. Similarly, in constructing multiple-choice or other objective type tests of observational and interpretive skills, it is possible to design each question to include (in addition to the correct answer) a series of alternatives representing each major type of error and to score these according to such categories as the following:

1. Failure to observe (i.e., the percent of significant findings not noted),
2. Overinterpretation (i.e., the percent of answers that go beyond the data),
3. Overcaution (i.e., the percent of answers that do not exploit the data fully), and

4. Crude errors (i.e., the percent of answers interpreting normal findings as pathological and vice versa).

Clinical Judgment

To assess judgment, which is certainly one of the most important aspects of clinical competence, various types of exercises have been developed to simulate the physician-patient encounter. These simulated problems in patient management all have the following characteristics: First, each exercise is initiated by information of the type a patient gives a physician, not by a pre-digested summary of the salient features of the case. To be realistic, this information is couched in terms which would be used by the patient or a referring physician. Second, the exercise is designed to require a series of sequential interdependent decisions representing the various stages in the diagnostic work-up and management of the patient whose presenting complaints have been described in the introduction. Third, the examinee is able to obtain information in realistic form about the results of every decision which he or she makes; this information is then available as a basis for subsequent action. Fourth, a special chemical is used in printing the answer sheet so that once these data have been obtained, it is impossible for the examinee to retract a decision that is revealed to be ineffectual or harmful. Fifth, the problem is constructed to allow for both different medical approaches and variation in patient responses appropriate to these several approaches. Sixth, provision is made for modifications in the problem and in the simulated patient's responses, depending on the specific interventions chosen by each examinee. Finally, these modifications differ among examinees according to the unique configuration of prior decisions each has made.

It is this last characteristic which is of greatest significance and which most clearly differentiates the resulting exercises from the Tab Test of Glaser and his associates (Glaser, Damrin, & Gardner, 1960), the Diagnostic Skills Test of Rimoldi (1963), and the National Board Part III Examination for Interns (Hubbard, 1964), all of which have served as a point of departure for our work. However, these earlier tests are in linear form: each examinee is confonted with the same problem, which remains identical throughout for all respondents. Thus a premium is necessarily placed either on the efficiency of reaching a single correct solution or on the appropriateness of each decision considered independently. In contrast, the branching problems (McGuire, Solomon, & Bashook, 1976) which are presently being used require the examinee to make revealing choices from an almost unlimited number of broad strategic routes, several of which may lead to acceptable results.

Problems of this type have now been developed in three modalities for use with medical students, residents in training, physicians in practice, nurses, and allied health personnel. The three modalities are as follows (1) A paper-and-pencil format using latent image or opaque overlay techniques for feedback systems in problems suitable for either individual or group administration and amenable to computer scoring and analysis (McGuire, Solomon, & Bashook, 1976); (2) A computer format employing unconstrained natural language suitable for individualized administration to untrained subjects (Harless et al., 1971); and (3) an oral interactive format suitable for role-playing exercises in individual or small group settings (Levine & McGuire, 1970).

In the exercises developed for medical personnel (McGuire, Solomon, & Forman, 1976), each problem is initiated by a brief verbal description of the patient's chief complaint or by a short color film in which the patient describes his or her current illness. The examinee must then decide how to approach this patient (i.e., what, if any, further work-up seems indicated at this point). That decision is recorded either by erasing the opaque overlay or by using a chemical marker to develop the latent image on a specially constructed answer sheet. The act of recording a decision reveals an instruction directing the examinee to a section of the booklet appropriate to his or her choice; the section contains a long list of inquiries or interventions, each of which (if chosen) evokes a response from the patient.

Each problem contains many such sections; not all of them are necessarily relevant to the optimal management of the patient. The sections are arranged in scrambled order and may be sealed to minimize the possibility of using the options offered in them as clues to the expected behavior. In each new section the examinees must indicate their choices from a series of specific actions, and at each stage they must make a strategic decision about the next step in the overall management of the patient. This decision determines the section to which the examinee is next directed. In this fashion a problem may be carried through many stages, and at each stage the examinee must make further decisions based on the specific responses of the patient (which had been evoked by the examinee's earlier inquiries and interventions).

The stages in the work-up and the responses to the specific procedures the examinee may elect are meticulously designed to simulate an actual clinical situation. Results of diagnostic and therapeutic maneuvers are reported in a form resembling that which the physician is accustomed to encounter. For example, in response to an order for a specific test, a laboratory report is revealed. In response to an order for an X-ray, electroencephalogram, or electrocardiogram, the examinee is referred to a high quality photographic reproduction of the X-ray or tracing;and in response to an order for a blood smear, a color plate of the smear is provided. In response to a request for auscultatory data,the candidate may be referred to a high fidelity tape recording of heart, breath, or abdominal sounds reproduced through individual stethophones. In response to medication prescribed,the patient's reaction is reported. No interpretation of these data is offered and none is explicitly demanded. If the student requests a consultation, appropriate help is forthcoming. Otherwise, the examinee is simply given the data he or she requests and is required to act on those data as does the physician in the conventional clinical setting.

The complications which must be managed differ from candidate to candidate,depending (as they do in the office or clinic) on the unique combination of specific procedures each has elected at earlier stages. For some the answer sheet will reveal an instruction to skip one or more sections of a problem entirely because the approach they have chosen is effective in avoiding potential complications. If, however, at any stage the examinee orders something harmful or fails to take measures essential to the recovery of the patient, a description of the clinical features of the complication that has developed is provided. The examinee is then directed to a special section in which opportunity is provided to take heroic measures to rectify previous errors. If the remedial measures chosen at this point are inadequate, the examinee may be instructed that the problem is terminated because the patient has suffered a relapse, gone to another physician, or expired.

Various users have developed a variety of systems for scoring clinical simulation, depending on the nature of the problems and the format in which they are presented. The most commonly used scores are:

1. Proficiency (i.e., percent agreement with expert consensus),
2. Efficiency (i.e., percent of total decisions that are helpful),
3. Competence (i.e., a weighted product of proficiency and efficiency),
4. Errors of omission (i.e., the percent of helpful inquiries and interventions not chosen),
5. Errors of commission (i.e., the percent of harmful inquiries and interventions chosen).

This set of scores can be obtained on a single simulation, a set of several simulations, or on different elements (e.g., data gathering, management strategy, therapy) in one or more simulations.

Interpersonal Skills

The physician's interpersonal skills and attitudes can be assessed by simulated interviews and conferences in which an individual--not necessarily an actor--is programmed to take the part of a patient, a colleague, or other member of the health team. These simulations differ from the live simulations of clinical problems only in their focus. In the former, the emphasis is on solving a complex medical problem; in the latter, the situations are chosen to focus on a physician's communication skills, sensitivity to patient and colleague needs, and attitudes toward these important people.

For example, in what purports to be a straight-forward diagnostic interview, the simulated patient may (if appropriately handled by the physician) provide cues that the real problem is a deep-seated personal, familial, or emotional one. Readily assayed in such simulated interviews are the physician's sensitivity and response to these cues, skill in communicating with and managing the patient, and skill in gaining the patient's understanding and cooperation in a proposed plan of management (Levine & McGuire, 1968, 1970).

Analogous simulated interviews with colleagues have been developed to deal with referral and consultation requests and even with those ticklish situations in which colleagues differ sharply about the management of a particular patient. Clearly, in such settings a physician's skills and professional attitudes are soon apparent. The simulated interview may also be designed so that the physician is required to communicate with one or more members of the health team. For example, he or she may be required to give instructions to a simulated nurse, to request assistance for the patient from a simulated dietitian or social worker, or to make a presentation and respond to the reactions of other members of the simulated health team in a staff conference. The candidate can be objectively rated on relevant communication skills, interpersonal skills, attitudes, and professional demeanor sampled in such situations.

Technical Skills

In order to test the physicians' technical skills in a simulated setting, it is necessary to have access either to a very large number of remunerated patients with stable conditions who present varied pathology or to three-dimensional models simulating different pathologies in a life-like form which can be examined by the physician (Penta & Kofman, 1963). Several such models (e.g., the Colenbrander Opthalmascopic Manikin, a laryngeal model, a lymph node simulator, heart or breath sound simulators) are now available, each with a variety of interchangable pathologies. Using these three-dimensional and electronic simulators, students can be required to demonstrate that they have mastered the skills of using sophisticated instruments (e.g., the opthalmascope, the otoscope) to collect data and to interpret them. While offering great promise as a means of objectively testing technical skills at all levels, most of the presently available models are still too crude for use with advanced students or practicing physicians.

In a few areas computerized robots and live actors (programmed to present certain findings) may be effectively employed. In using these simulators (i.e., models, robots, and actors) the student may be observed and rated on technique and/or may simply be asked to examine a series of simulators that present varied pathology. The student then records the findings in a manner permitting mechanical scoring at a later time. For purposes of diagnostic testing and educational guidance, both technique and accuracy are rated; for purposes of certification only the latter may be necessary.

Conclusions

It is generally accepted that the best predictor of future performance is past performance of a closely similar type. Implementation of this principle would seem to dictate the use of such methods as work samples and assessment of on-the-job performance as the methods of choice in evaluating occupational and professional competence. However, in some occupations and professions such techniques are unfeasible; in others their employment would cause the examinee or others to be subjected to unacceptable hazards. Furthermore, even where observation of on-the-job performance is possible, it may be highly unreliable because the work samples cannot be standardized, are necessarily limited in number and variety, and are subject to situational constraints making generalization difficult. Thus, in most occupations and professions simulation techniques will yield more valid, reliable, and representative performance data than that obtained from either real life observation or more conventional methods of testing.

References

Glaser, R., Damrin, D., & Gardner, F.M. The tab item: a technique for the measurement of proficiency in diagnostic problem-solving tasks. In A. A. Lumsdaine & R. Glaser (Eds.), Teaching Machines and Programmed Learning: A Source Book. Washington DC: National Education Association, 1960.

Harless, W., Drennan, G., Marxer, J., Root, J., & Miller, G. CASE: A computer-aided simulation of the clinical encounter. Journal of Medical Education, 1971, 46, 443-48.

Hubbard, J.P. Programmed testing. In Examinations and Their Role in Evaluation of Medical Education and Qualification for Practice. Philadelphia: National Board of Medical Examiners, 1964.

Levine, H.G., & McGuire, C.H. The validity and reliability of oral examinations in assessing cognitive skills in medicine. Journal of Educational Measurement, 1970, 7, 63-74.

Levine, H.G., & McGuire , C.H. Role playing as an evaluative technique. Journal of Educational Measurement, 1968, 5, 1-8.

Levine, H.G. & McGuire, C.H. The use of role-playing to evaluate affective skills in medicine. Journal of Medical Education, 1970, 45, 700-705.

McGuire, C., Solomon, L., & Forman, P. Clinical Simulations: Selected Problems in Patient Management (2nd ed.). New York: Appleton-Century-Crofts, 1976.

McGuire, C., Solomon, L., & Bashook, P. Construction and Use of Written Simulations. New York: The Psychological Corporation, 1976.

Penta, F.B., & Kofman, S. The effectiveness of simulation devices in teaching selected skills of physical diagnosis. Journal of Medical Education, 1973, 48, 442-45.

Rimoldi, H.J.A. Rationale and application of the test of diagnostic skills. Journal of Medical Education, 1963, 38, 364-73.

Acknowledgments

Portions of this paper were adapted from McGuire, Solomon, & Bashook, 1976.

PROBLEMS OF PERFORMANCE MEASUREMENT IN COMPUTER-BASED SIMULATION

BRUCE W. KNERR
U.S. ARMY RESEARCH INSTITUTE

A major theme in research in training and education today is individualization. Underlying this theme is the belief that the instructional process, including both the delivery and the assessment of instruction, can be conducted more efficiently or more effectively if the process is adapted to the characteristics and performance of the individual student. Computerized adaptive testing is one method of providing such individualization in the area of assessment. In the area of instructional delivery, there are a variety of methods of providing individualization, encompassing a wide range of complexity.

At the simplest extreme, the training content is held constant for all students, but each student progresses at his/her own rate. Testing is required at the end of each unit of instruction, with the students who fail the test repeating that unit. This type of individualization can be used with a variety of instructional methods and media, ranging from a linear programmed text to performance-based (hands-on) training.

At the opposite extreme, perhaps the most complex form of individualization is represented by a mixed-initiative dialogue between a student and a "private tutor." In this situation students can ask questions of the tutor, to which the tutor will respond, and the tutor can question students in order to assess their proficiency. If the tutor is a human expert, the cost of this type of training can be quite high. Recently, the use of "artificial intelligence" techniques has made computer-based tutors possible. SOPHIE (Brown, Burton, & Bell, 1974) and SCHOLAR (Collins, Warnock, & Passafiume, 1974) are examples of two systems which use computer-based tutors. The sophisticated computer support required for the operation of such systems, however, tends to prevent their use in a "real world" training environment.

Between these two extremes lie a variety of methods for individualizing training. The use of computers has provided increased capabilities for the individualization of training; yet these capabilities are often under-utilized, at least within the Army. One reason for this is the time, and consequently the cost, required to prepare complex instructional materials or alternate sets of instructional materials. The more complex the individualization provided, the more time-consuming and costly the lesson material development becomes. In operational settings, where computer-based training must be implemented in a limited period of time, individualization is often sacrificed for efficiency of production.

In an effort to develop methods for simplifying the process of preparing individualized training materials, the Army Research Institute (ARI) initiated a program in 1974 to investigate the feasibility of using methods derived from artificial intelligence techniques for Army training. The content area selected was electronic troubleshooting training, an area which seemed readily amenable to the type of training approach proposed and, in addition, represented a training problem.

Current Army troubleshooting training is equipment specific and hands-on. The student learns the sequential procedures necessary to troubleshoot a specific item of equipment, then practices on the equipment itself. This method has several advantages. It insures that the student has mastered certain prerequisite skills, such as the use of test equipment. It teaches the student the physical layout of the equipment and the correspondence between the equipment and the circuit diagram. Finally, it gives the student practice in the assembly and disassembly of the equipment.

There are also several disadvantages. Since the training is equipment specific, there is little transfer to other items of equipment. A substantial amount of equipment is required for training purposes. Instructors must spend large portions of their time inserting malfunctions into the equipment, rather than actually conducting training. Much student time is spent assembling, disassembling, and soldering the equipment; this reduces the number of different problems students can experience during their training.

The Adaptive Computerized Training System

The system that has been developed under this effort is called the Adaptive Computerized Training System (ACTS). All work accomplished through January 1977 was performed by Perceptronics, Inc., under contract to ARI. Details are contained in four reports: May, Crooks, Purcell, Lucaccini, Freedy, and Weltman (1974); Kuppin (1976); May, Crooks, and Freedy (1976); and Crooks, Kuppin, and Freedy (1977).

When training with the ACTS, the student's task is to troubleshoot a complex electronic circuit by making various test measurements, replacing the malfunctioning part, and making final verification measurements. All equipment is simulated by the ACTS. The heart of the system is an adaptive computer program which "learns" the student's decision utility structure, compares this structure to that of an expert, and when complete, will adapt the instructional sequence and feedback to eliminate discrepancies between the two. An Expected Utility (EU) model of decision-making is the basis of the models of both the student and expert trouble-shooters.

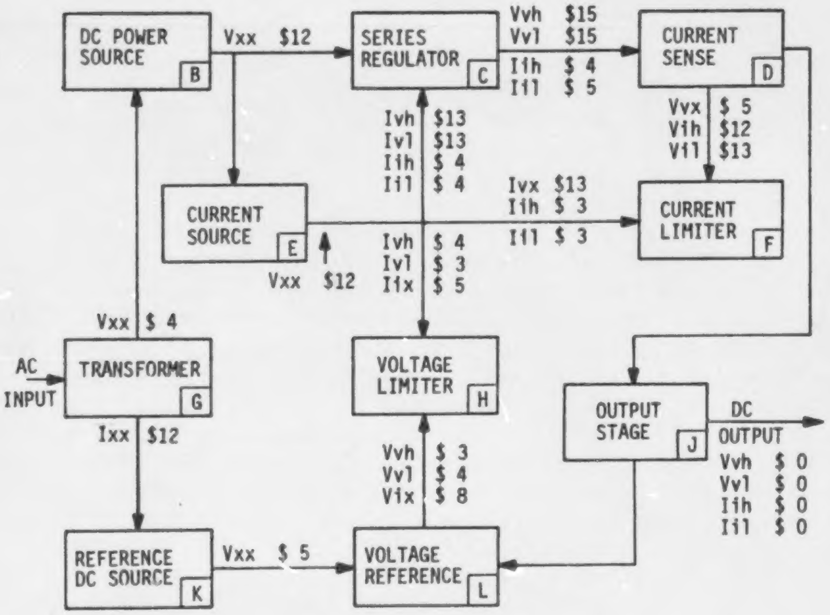
The ACTS is not being proposed as a complete troubleshooting training method. It will not train students to use test equipment, assemble or disassemble the equipment, nor provide them with background information about the operation of the circuit. It is designed to train students in the decision-making aspect of the troubleshooting process.

ACTS Components

The ACTS consists of four major components: (1) the task model; (2) the expert model; (3) the student model; and (4) the instructional model.

Task model. The task model is a simulation of the circuit on which the student is to be trained. The circuit currently being used is a modular version of the Heathkit IP-28 Power Supply.[1] A simplified schematic diagram of this circuit is shown in Figure 1. The power supply, when functioning properly, converts a 117-volt alternating current input (shown at left) into a stable, low voltage, low amperage direct current output (shown at the right). As the diagram indicates, the circuit consists of ten modules. Since the output of the circuit must be stable, even with variations in the input, there are a number of corrective feedback loops in the circuit which make the troubleshooting process more difficult.

Figure 1
ACTS Circuit Diagram
(from Crooks, Kuppin, & Freedy, 1977)


DC POWER SOURCE B
Vxx $12
SERIES REGULATOR C
Vvh $15
Vvl $15
Iih $ 4
Iil $ 5
CURRENT SENSE D
Ivh $13
Ivl $13
Iih $ 4
Iil $ 4
Vvx $ 5
Vih $12
Vil $13
CURRENT SOURCE E
Ivx $13
Iih $ 3
Iil $ 3
CURRENT LIMITER F
Ivh $ 4
Ivl $ 3
Iix $ 5
Vxx $12
Vxx $ 4
AC INPUT
TRANSFORMER G
VOLTAGE LIMITER H
OUTPUT STAGE J
DC OUTPUT
Vvh $ 0
Vvl $ 0
Iih $ 0
Iil $ 0
Ixx $12
Vvh $ 3
Vvl $ 4
Vix $ 8
REFERENCE DC SOURCE K
Vxx $ 5
VOLTAGE REFERENCE L

Rather than using a complete simulation of the circuit, as does SOPHIE (Brown et al., 1974), the ACTS uses a table-driven simulation of the results of faults in the circuit. That is, once a fault has been selected, the ACTS reads the appropriate measurement outcomes from a preprogrammed table.

---

[1] Commercial designations are used only for precision of description. Their use does not constitute endorsement by the Department of the Army or the Army Research Institute.

Expert model. The second major component of the ACTS is a model of an expert troubleshooter. This is an EU decision model which predicts the expert's measurement choices as he/she troubleshoots the circuit. It is developed through on-line observation of the expert's troubleshooting behavior.

Student model. The student model, like the expert model, is an EU decision model which predicts the student's measurement choices. It is developed through on-line observation of the student's behavior as he/she solves troubleshooting problems on the ACTS.

Instructional model. The last major component is the instructional model. The function of the instructional model is to compare the expert and student models, determine discrepancies between the two, and modify the instructional feedback and problem presentation sequence in order to reduce those discrepancies. Currently, the instructional model can provide some adaptive feedback, but cannot modify the instructional sequence.

The Expected Utility Model

The expert and student models are central to operation of ACTS. While the two models serve different functions and use different data, their operation is identical.

Consider an expert troubleshooter who is given a defective IP-28 power supply and simply is told, "It doesn't work. Fix it." If he/she chooses to repair it, there are a limited number of actions that can be taken. The switches and meters on the front panel of the power supply can be used to check any of the four outputs, any of the 25 possible internal test measurements can be taken, or any of the ten circuit modules can be replaced.

Each of these possible actions has associated with it a set of possible outcomes. For example, measuring the output voltage with the voltage setting in the high state could produce outcomes of normal, low, very low, or zero.[2] Module E, the current source, could be good (in which case replacing it would not correct the circuit malfunction) or bad (in which case replacing it would correct the malfunction).[3]

Each possible outcome has three properties. The first is the conditional probability of the occurrence of that outcome, given that the appropriate action is selected and given the measurement outcomes previously obtained. The second property is the utility of the outcome to the troubleshooter, i.e., what he/she gains or loses as a result of the outcome. Utility is subjective, but it should be related to the cost (in time or money) of taking that action.

---

[2] When using the actual equipment, of course, the outcome would be a numeric value. The troubleshooter would next have to determine whether this value was high, medium , low, etc.

[3] Again, this is a simplification. There are a number of ways in which the current source could be "bad."

The third property is the amount of information that the outcome provides about the location of the fault. These properties are combined as follows:

$$EU_j = \sum_i^n \alpha_{ij} P_{ij} U_{ij} \quad , \qquad [1]$$

where

EU_j = the expected utility of action A_j;
P_{ij} = the probability that outcome i, of a set of n outcomes will occur if action A_j is selected;
U_{ij} = the utility of outcome i of action A_j; and
α_{ij} = the information gain resulting from the occurrence of outcome i of action A_j.

The model development process starts by obtaining the probabilities required by the model. Initially, two different sets of probabilities are needed, and the others can be obtained from them. The first set of probabilities is $P(K)$, the a priori probability of occurrence of each fault, K. There are several ways to obtain $P(K)$: (1) examination of maintenance records to determine the relative frequency of each fault; (2) expert opinion; or (3) assuming that all faults are equally likely. The second set of probabilities needed is $P(X_{ij}|K)$, the probability of obtaining outcome i of measurement j, given fault K. This set of probabilities can be obtained from an analysis of the circuit or from expert opinion.

Given these two sets of probabilities, $P(X_{ij}|\overline{X})$, the probability of a particular outcome, given the entire measurement history, can be calculated by the formula

$$P(X_{ij}|\overline{X}) = \frac{\sum_{k \epsilon A} P_k}{\sum_{k \epsilon B} P_k} \quad , \qquad [2]$$

where $A = L_{ij} \cap L_x$;
$B = L_x$;
L_{ij} = the set of faults possible given outcome X_{ij}; and
L_x = the set of faults still possible, given the measurement history, $\overline{X}$.

In other words, the probability of an outcome, given the previous measurement history, is the sum of the probabilities of the faults associated with that outcome *and* not eliminated, divided by the sum of the probabilities of all faults not eliminated. As measurement outcomes are obtained, the $P(K)$ are updated to reflect the changing measurement history.

Since each circuit module contains more than one possible fault, it is also necessary to calculate the probability of a particular module being bad. This is simply the sum of the probabilities of all possible faults within that module.

The information gain component, α_{ij}, also needs to be calculated before the expert utilities can be estimated. While alternative methods of calculation are available, the formula currently used is

$$\alpha_{ij} = -f_i \log_2 f_i \quad , \qquad [3]$$

where

$$f_i = \frac{\text{Number of possible faults remaining associated with outcome } i}{\text{Number of possible faults remaining}} \quad .$$

The α_{ij} are calculated directly. Referring once again to the expected utility formula (Equation 1), it can be seen that there are now two sets of unknowns remaining: The expected utilities for each possible action (EU_j) and the utilities for each possible outcome (U_{ij}). These unknowns are estimated by "tracking" an expert's troubleshooting behavior as he/she completes a series of problems on the ACTS. As these are completed, the expert is presented with the updated probabilities of measurement outcomes. The values of the known model parameters are entered into the expert model before the expert starts, with the utilities (U_{ij}) set at some common arbitrary value (usually 100). The expert model chooses the action which has the highest expected utility. If the expert then chooses the same action, no changes in the model are made.

However, if the action selected by the expert model differs from the action selected by the expert, the model utilities associated with the model choice are punished (decreased) and those associated with the expert choice are rewarded (increased). These processes are described by the following formulae:

$$U_{ij}^{t+1} = U_{ij}^{t} - P_{ij}\,\alpha_{ij}\,\gamma \qquad \text{(Punish)} \qquad [4]$$

$$U_{ij}^{t+1} = U_{ij}^{t} + P_{ij}\,\alpha_{ij}\,\gamma \qquad \text{(Reward)} \quad , \qquad [5]$$

where

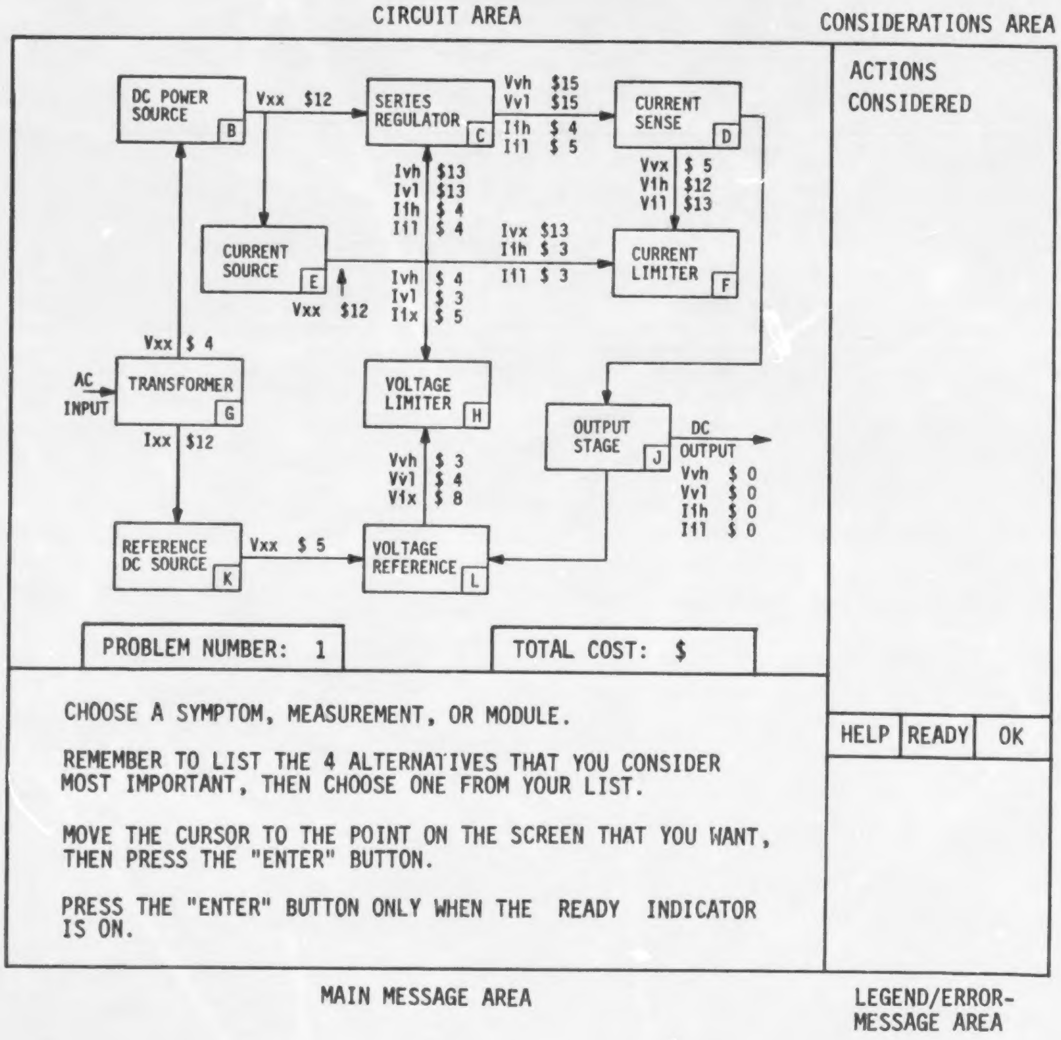
U_{ij}^{t} = the (unadjusted) utility of result i of action j at time t;
U_{ij}^{t+1} = the (adjusted) utility of result i of action j at time t+1;
γ = a constant;
P_{ij} = the probability of occurrence of result i of action j; and

α_{ij} = the information gain resulting from result i of action j.

This process continues until the estimated utilities become stable. This will occur when the expert model is able to predict the choices of the expert accurately.

At this point the expert is no longer needed. The expert model, having been "trained," replaces the human expert. Now the system is ready to begin training the student. As the expert did previously, the student now begins to troubleshoot. As students do this, they have access to the probabilities produced by the expert model.

Figure 2
The ACTS Student Display as it
Appears at the Start of a Problem


CIRCUIT AREA
CONSIDERATIONS AREA
DC POWER SOURCE B
Vxx $12
SERIES REGULATOR C
Vvh $15
Vvl $15
Iih $ 4
Iil $ 5
CURRENT SENSE D
Ivh $13
Ivl $13
Iih $ 4
Iil $ 4
Vvx $ 5
Vih $12
Vil $13
CURRENT SOURCE E
Ivx $13
Iih $ 3
Iil $ 3
CURRENT LIMITER F
Vxx $12
Ivh $ 4
Ivl $ 3
Iix $ 5
Vxx $ 4
AC INPUT
TRANSFORMER G
VOLTAGE LIMITER H
OUTPUT STAGE J
DC OUTPUT
Vvh $ 0
Vvl $ 0
Iih $ 0
Iil $ 0
Ixx $12
Vvh $ 3
Vvl $ 4
Vix $ 8
REFERENCE DC SOURCE K
Vxx $ 5
VOLTAGE REFERENCE L
PROBLEM NUMBER: 1
TOTAL COST: $
ACTIONS CONSIDERED
HELP
READY
OK
MAIN MESSAGE AREA
LEGEND/ERROR-MESSAGE AREA

CHOOSE A SYMPTOM, MEASUREMENT, OR MODULE.

REMEMBER TO LIST THE 4 ALTERNATIVES THAT YOU CONSIDER MOST IMPORTANT, THEN CHOOSE ONE FROM YOUR LIST.

MOVE THE CURSOR TO THE POINT ON THE SCREEN THAT YOU WANT, THEN PRESS THE "ENTER" BUTTON.

PRESS THE "ENTER" BUTTON ONLY WHEN THE READY INDICATOR IS ON.

As the student solves a series of problems, the student model, functioning in the same manner as the expert model, learns the student's utilities. When the estimated student utilities begin to stabilize, feedback can be provided to the student.

Student Interactions

The student display as it appears at the start of a problem is shown in Figure 2. It has four areas: (1) the circuit area; (2) the main message area; (3) the considerations area; and (4) the legend/error-message area. The circuit area contains a diagram of the circuit, along with the measurements that the student can take and the cost for each measurement.[4] After a measurement has been made, the outcome is displayed instead of the cost. The considerations area is used to present the outcome probabilities to the student. The main message area informs the student of the options available and also provides instructions and assistance. The legend/error-message area provides a legend interpreting the codes by which the probabilities are displayed and is also used to inform students when they have taken an "illegal" action. All student inputs are accomplished through the use of a track-ball and cursor.

Figure 3 presents a flow diagram of the student interactions with the ACTS during the training process. At the start of a problem, students are told only that the circuit has a malfunction which they are to correct. Initially, the student can select one of five options. The first is to ask for HELP. As shown in Figure 3, this option is available to students whenever they are required to select some action. HELP provides the student with two types of information from the expert model: (1) a list of the circuit modules that could be bad; and (2) the action that the expert would take next.

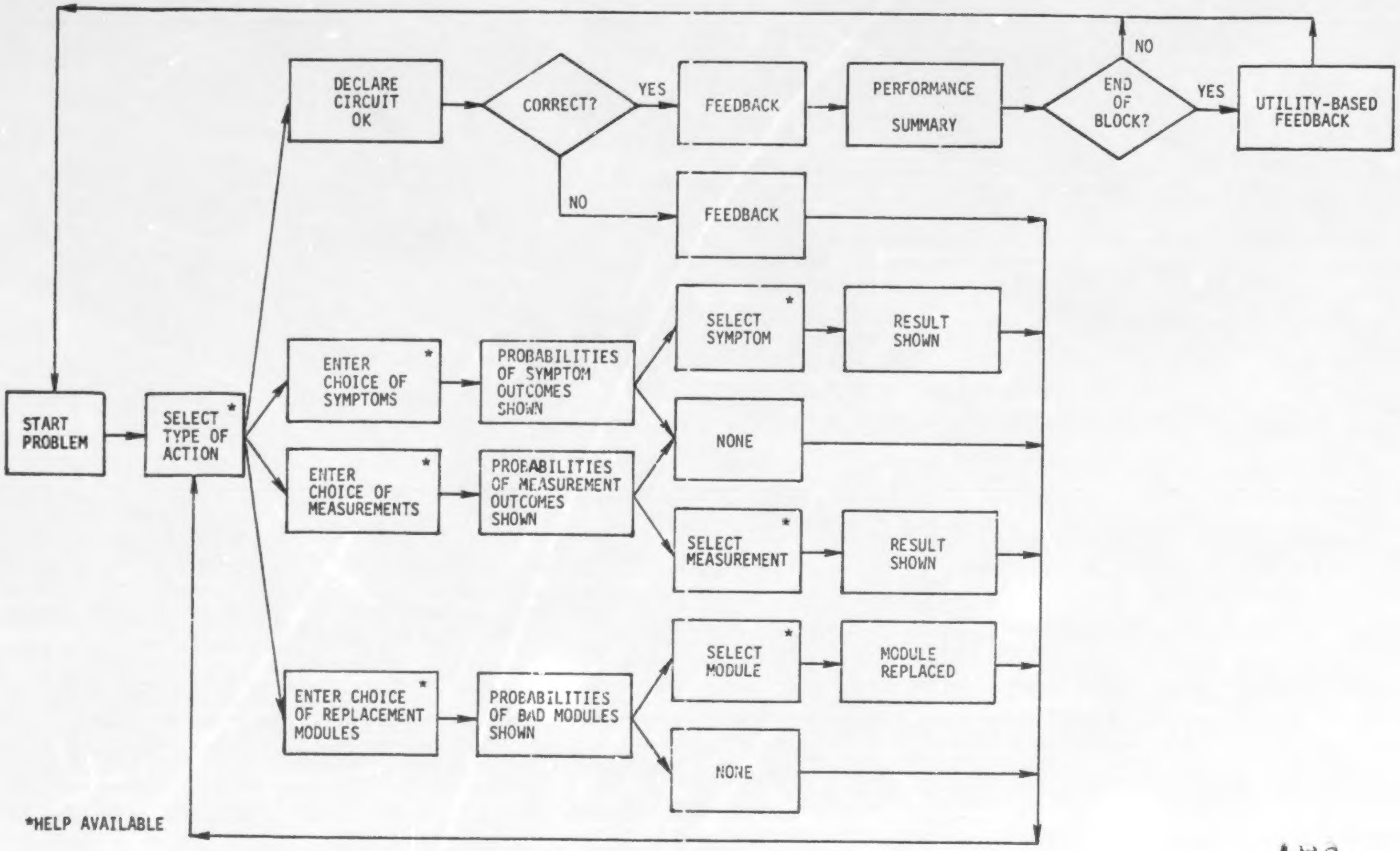
The second option that students can select is to enter their choice of symptoms for consideration. Symptoms, which are measurements made on the final output of the circuit, are distinguished from the internal measurements. After the student chooses a symptom for consideration, the expert model's probabilities for the possible outcomes of a check of that symptom are shown. The student may choose to consider additional symptoms, select one of the symptoms considered, or choose none of the symptoms and return to the action selection choice point. HELP is also available. Selecting a symptom causes the outcome to be displayed and then returns the student to the action selection choice point.

The third option that students can select is to enter their choice of (internal) measurements for consideration. The sequence of interactions for this option is the same as the sequence for the previous option, entering a choice of symptoms for consideration.

---

[4] Costs are presented in dollar amounts. They reflect the amount of time that would be required to take the test measurement on the actual equipment.

Figure 3
Student Interaction Sequence


DECLARE CIRCUIT OK
CORRECT?
YES
NO
FEEDBACK
PERFORMANCE SUMMARY
END OF BLOCK?
NO
YES
UTILITY-BASED FEEDBACK
FEEDBACK
START PROBLEM
SELECT TYPE OF ACTION *
ENTER CHOICE OF SYMPTOMS *
PROBABILITIES OF SYMPTOM OUTCOMES SHOWN
SELECT SYMPTOM *
RESULT SHOWN
NONE
ENTER CHOICE OF MEASUREMENTS *
PROBABILITIES OF MEASUREMENT OUTCOMES SHOWN
SELECT MEASUREMENT *
RESULT SHOWN
ENTER CHOICE OF REPLACEMENT MODULES *
PROBABILITIES OF BAD MODULES SHOWN
SELECT MODULE *
MODULE REPLACED
NONE
*HELP AVAILABLE

The fourth option is to enter a choice of modules to be considered for replacement. After the student chooses a module for consideration, the expert model's probability that that module is defective is shown to the student. The student may consider additional modules, choose to replace a specific module, or return to the action selection choice point. Again, HELP is available. If the student chooses to replace a module, the replacement is simulated and the student is returned to the action selection choice point.

The student's final option is to declare the circuit OK. If the student is incorrect, feedback to this effect is given and the student is returned to the action selection choice point. If a correct answer is given, the student is informed that the circuit malfunction has been corrected. At this point the utilities in the student model are updated, and a performance summary is printed for the experimenter. If a block of problems has been completed, the student may be provided with feedback based on the student model utilities. Otherwise, a new problem is initiated.

Adapting Training

The ACTS has still not been completed. The major difficulty has been the development of the mechanism (or mechanisms) by which the training is to be adapted to individual performance. Ideally, such mechanisms should affect both the feedback that the student receives and the sequence in which problems are presented and should be based on the student's utilities. This has turned out to be more complex than anticipated. One reason is the sheer volume of information provided by the utilities. There is one utility for each measurement outcome and one for the replacement of each module. In the case of the IP-28 power supply, there are 96 different utilities. Table 1 shows a sample set of student utilities.

Existing Mechanisms

Thus far two methods of providing feedback to the student, one of which uses student utilities, have been developed. Neither method alters the problem presentation sequence. The first, and simplest, method presents the expert model's action choice to the student after the student has made a selection and obtained the result. Student utilities do not play a part in this feedback sequence, but the expert model is required to make it practical. Determination of the "best" action at any point in the troubleshooting sequence must consider the previous actions taken and results obtained by the student. Between 4 and 12 sequential actions (decisions) are required to repair the circuit; and initially, the student can choose from among 39 possible actions. To program all of the best choices using a logical branching technique would be, at best, a rather complex task.

The seoond method of providing feedback is based on comparisons among "key" student utilities. The key utilities are those for measurements identified by an expert as being of critical importance in the fault isolation process. For the IP-28 power supply, the utilities for the outcomes for six measurements were considered to be the key utilities. Based on the relationships among these utilities, a set of six decision rules and feedback

Table 1
A Sample Set of Student Utilities

Measurement or Symptom Code	Utility for Normal Outcome	Utility for Non-normal Outcome 1	Utility for Non-normal Outcome 2	Utility for Non-normal Outcome 3	Module Code	Utility for Replacement
P	100	100	100	100	B	100
Q	124	113	100	100	C	100
R	101	100	92		D	100
S	[illegible]	60	81		E	100
1	100	100	100		F	100
2	100	100	100		G	100
3	100	100	100		H	100
4	47	34	90		I	100
5	100	100	100		J	100
6	100	100	100		K	100
7	100	86	86		L	100
8	100	86	86			
9	100	100	100			
10	100	100	100			
11	100	100	100			
12	100	100	100			
13	100	100	100			
14	100	100	100			
15	100	100	100			
16	100	100				
17	100	113	113			
18	100	100	100			
19	100	100	100			
20	108	100	100			
21	100	100	100			
22	100	100				
23	100	100	100			
24	86	100	113			
25	100	100				

statements were developed. Samples are shown in Table 2. The decision rule for the first feedback statement should be read as follows:

> If any of the utilities for Measurement 3 are less than any of the utilities for Measurement 19, or if any of the utilities for Measurement 3 are less than any of the utilities for Measurement 11, present this feedback statement to the student.

The appropriate feedback statements are initially presented to the student after the fifteenth problem has been completed, with updated statements presented every fifteen problems thereafter. The student can review them at any time.

Table 2
Sample Decision Rules and Feedback Statements

Rule	Feedback
3<19 or 3<11	Although Measurement 3 is located at a good point to isolate the power input modules, it is expensive. Use this measurement after you have eliminated most other possibilities. Measurement 3 should be used when the probability of a normal outcome is rather high, but not certain (a range of 60% to 80%).
2<11 or 3<11 or 4<11	A good first step in checking the operation of current and voltage feedback loops is to check the output of the series regulator. This should be done with the circuit operating at full output since this fully exercises the circuit functions. Therefore, Measurement 9 or 11 should be used even if there is a low probability of a normal outcome. Use Measurement 11, since it is much cheaper than 9.

Alternate Methods

These approaches, while promising, do not use all of the information available about student performance. The first makes no use of student utilities, and the second uses only a subset of them. Possible alternate methods which use more of the information provided by the student utilities are currently under consideration. All of them rely on the difference between corresponding student expert and student utilities. The difference could be based on either outcomes (utilities) or actions, as shown in Equations 6 and 7:

$$D_{ij} = U_{sij} - U_{eij} \qquad [6]$$

$$D_{\cdot j} = \sum_i (U_{sij} - U_{eij}) \quad , \qquad [7]$$

where

D_{ij} = difference between student and expert utilities for outcome i of measurement j;
$D_{\cdot j}$ = difference between student and expert utilities for measurement j;
U_{sij} = student utility for outcome i of measurement j; and
U_{eij} = expert utility for outcome i of measurement j.

Whatever method is used to calculate this difference, the following procedures would then be used to individualize instruction. The difference having the largest absolute value would be selected. This identifies the measurement or outcome for which the discrepancy in utilities between student and expert is greatest. If the difference is positive, the student has a higher utility for that measurement or outcome than does the expert. A positive sign should result in two actions: (1) presentation of feedback designed to produce decreased use of that measurement or outcome, and (2) presentation of a new problem which does not require the use of that measurement or outcome. A negative sign, which indicates that the student has a lower utility than the expert, should also result in two actions: (1) presentation of feedback designed to produce increased use of that measurement or outcome, and (2) presentation of a new problem which requires the use of that measurement or outcome for proper solution. This process should continue until all differences are less than some specified value.

Discussion

Problems

These possible approaches to adapting instruction raise some problems which will direct a portion of the future ACTS research program. One problem area is related to the expert utilities. The ACTS method assumes that there is one "best" set of expert utilities (i.e., either that all experts have the same utilities or, if there are differences among the experts, that one set of utilities produces superior performance according to objective standards). If training is to be based on the differences between student and expert utilities, this assumption becomes much more critical than it has been in the past. Are there, for some circuits, multiple sets of equally effective utilities? If so, can some be more readily mastered by students than others?

It is proposed that training continue until all student utilities differ from those of the expert by less than some specified amount. The second problem area is concerned with how large this allowable difference should be. There are really several questions here. How similar must two sets of utilities be in order to produce identical performance over a series of problems? Since humans tend to be less than perfectly consistent decision makers, human performance and model predictions will not agree perfectly. It is perhaps more appropriate to ask how similar two sets of utilities must be before they would predict identical actions according to the EU model. Given that this question can be resolved, use of the difference between student and expert utilities as a measure of student proficiency and thus as a criterion for stopping training must be considered. It is unlikely that students would be trained until they perform exactly like an expert; it is more likely that students would be trained only until they meet some lesser standard. Can utilities be used to define this standard, or must more traditional measures (e.g., cost per problem, decisions per problem) be used?

A third group of problems is related to the number of utilities to be used in the training process. The existing method for providing feedback employing student utilities uses only a subset--the utilities for those measurements which have been identified by an expert as being the most important. This simplified the process of providing feedback, making it contingent upon only 18 utilities for 6 measurements, rather than 86 utilities for 29 measurements and symptoms in the entire set of utilities. However, when only 18 utilities are used, some of the information about student (and expert) performance is lost. This information loss may or may not be critical in terms of the training effectiveness of the ACTS. What methods, subjective or objective, could be employed to reduce the number of utilities used and still produce an acceptable level of training effectiveness? Expert judgment represents a subjective method which has the advantage of determining the reduced utility set for new circuits without the necessity for collecting data on student performance. But its effectiveness and the extent to which experts agree on which measurements are the "key" measurements have yet to be determined; objective methods still need to be identified.

A fourth and final problem area is related to the development of higher-level diagnostics based on the student utilities. Do certain patterns of student utilities correspond to student deficiencies, such as a lack of understanding of transistor operation? If so, can these diagnostics be used to modify the training process?

Future Directions

It is planned that future ACTS research and development will follow two divergent paths. This is necessary because the resolution of the problems described above and the evaluation of the ACTS in an operational setting in the near future are both equally important goals. The operational evaluation path will begin with laboratory evaluations of the training effectiveness of the current version of ACTS, including an evaluation of the transfer of ACTS training to the real equipment. If this phase is successful, it will be followed by the installation of the ACTS at an Army school, where its effectiveness will be evaluated in an ongoing course of instruction. Simultaneously, research on the problems described above will be conducted with the goal of incorporating the results in a second-generation ACTS.

References

Brown, J. S., Burton, R. R., & Bell, A. G. SOPHIE: A sophisticated instructional environment for teaching electronic troubleshooting: An example of AI in CAI (BBN Report No. 2790). Cambridge, MA: Bolt, Beranek, & Newmann, Inc., March 1974.

Collins, A., Warnock, E. H., & Passafiume, J. J. Analysis and synthesis of tutorial dialogues (BBN Report No. 2789). Cambridge, MA: Bolt, Beranek, & Newmann, Inc., March 1974.

Crooks, W. H., Kuppin, M. A., & Freedy, A. Applicaton of adaptive decision aiding systems to computer-assisted instruction: Adaptive Computerized Training System (ACTS) (Perceptronics Technical Report PATR-1028-77-1). Woodland Hills, CA: Perceptronics, Inc., March 1976.

Kuppin, M. A. Adaptive Computerized Training System (ACTS): User's manual (Perceptronics System Doc. 1028-76-1). Woodland Hills, CA: Perceptronics, Inc., December 1976.

May, D. M., Crooks, W. H., & Freedy, A. Application of adaptive decision aiding systems to computer-assisted instruction: Experimental studies (Perceptronics Technical Report PATR-1023-76-1). Woodland Hills, CA: Perceptronics, Inc., March 1976.

May, D. M., Crocks, W. H., Purcell, D. D., Lucaccini, L. F., Freedy, A., & Weltman, G. Application of decision aiding systems to computer-assisted instruction (Perceptronics Technical Report PATR-1019-74-12). Woodland Hills, CA: Perceptronics, Inc., December 1974.

Acknowledgments

The findings in this paper are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

THE DECISION MEASUREMENT SYSTEM AS A MEANS OF TESTING PERFORMANCE BY SIMULATION

MYRON A. ROBINSON
DATA-DESIGN LABORATORIES

C. LEE WALKER
CENTRAL TEST SITE FOR PTEP
NAVAL GUIDED MISSILES SCHOOL

The Decision Measurement System (DMS) is a paper-and-pencil simulation of the operation or maintenance of the complex electronic and electromechanical equipment found on the Fleet Ballistic Missile (FBM) Submarines. The DMS was developed by Data-Design Laboratories under the sponsorship of the Navy's Strategic Systems Project Office and has been incorporated into the Personnel and Training Evaluation Program (PTEP), which is coordinated by the Central Test Site, Naval Guided Missiles School, Dam Neck, Virginia.

The term "performance," in the context of this report, consists of two aspects: perceptual-motor and analytical (cognitive). In maintaining the electronic and electromechanical equipment found on the FBM submarines, the perceptual-motor aspect, which is confined to the use of test equipment and the removal and replacement of plug-in units is minimal. The analytical aspect predominates. In PTEP the DMS is used to assess the ability of technicians to follow established maintenance procedures, as well as to find innovative solutions to problems for which established procedures do not exist.

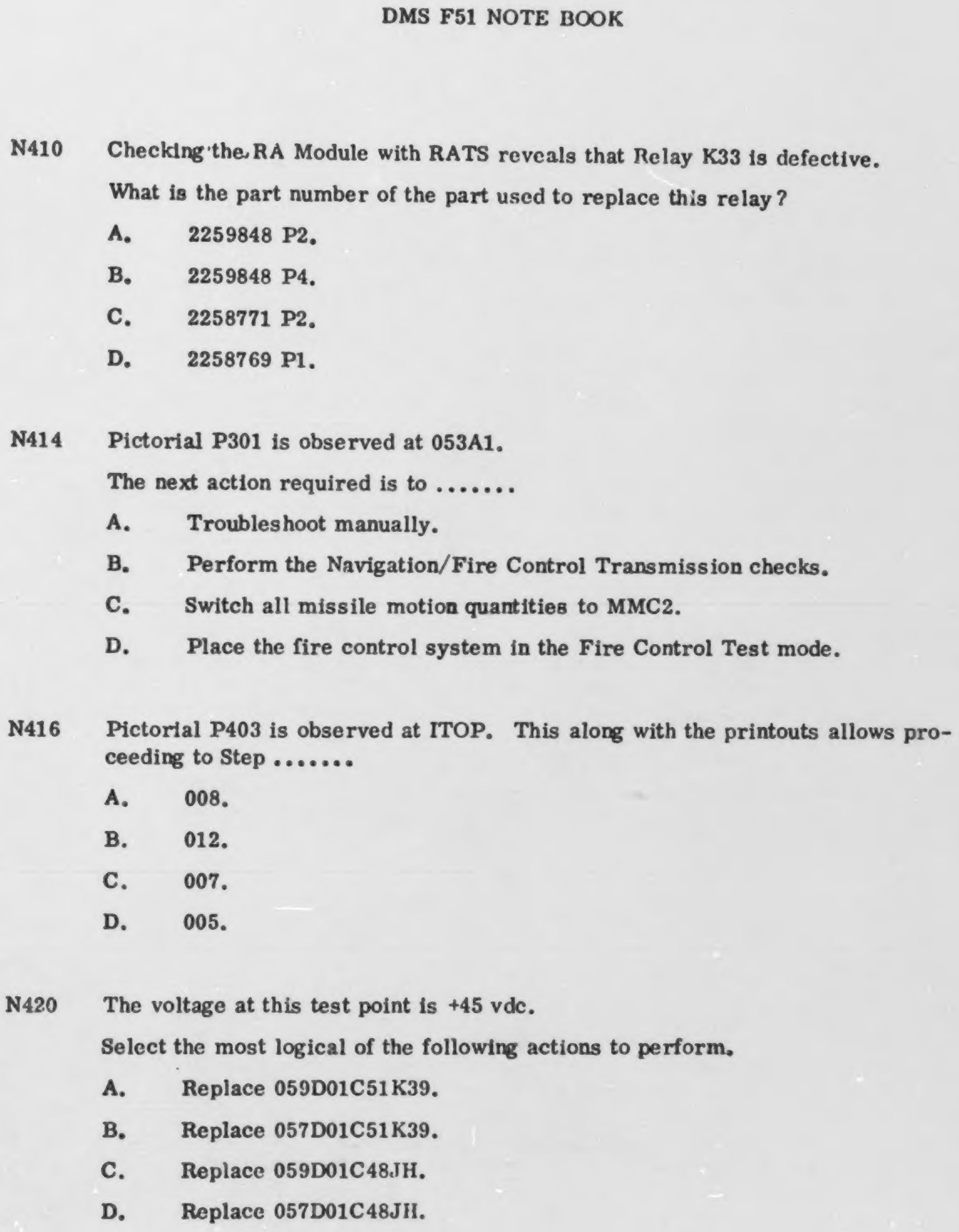
Description of the DMS

Format

A DMS Set consists of a Note Book, a Data Book, and an Exercise Sheet. The Note Book contains the set of test items (blocks) as shown in Figure 1. The items in the Note Book are scrambled, and the branching between items is controlled by the Exercise Sheet.

The Data Book contains the panel displays, wave forms, and diagnostic printouts referenced by the items in the Note Book (Figures 2, 3, and 4). For example, as shown in Figure 1, item N416 in the Note Book references Pictorial P403 in the Data Book. The examinee is directed to turn to the

Figure 1
Example of Page from Note Book



DMS F51 NOTE BOOK

N410 Checking the RA Module with RATS reveals that Relay K33 is defective.

What is the part number of the part used to replace this relay?

A. 2259848 P2.

B. 2259848 P4.

C. 2258771 P2.

D. 2258769 P1.

N414 Pictorial P301 is observed at 053A1.

The next action required is to ....

A. Troubleshoot manually.

B. Perform the Navigation/Fire Control Transmission checks.

C. Switch all missile motion quantities to MMC2.

D. Place the fire control system in the Fire Control Test mode.

N416 Pictorial P403 is observed at ITOP. This along with the printouts allows proceeding to Step ....

A. 008.

B. 012.

C. 007.

D. 005.

N420 The voltage at this test point is +45 vdc.

Select the most logical of the following actions to perform.

A. Replace 059D01C51K39.

B. Replace 057D01C51K39.

C. Replace 059D01C48JH.

D. Replace 057D01C48JH.

Figure 2
Example of Printout Page from Data Book

DMS F51 DATA BOOK

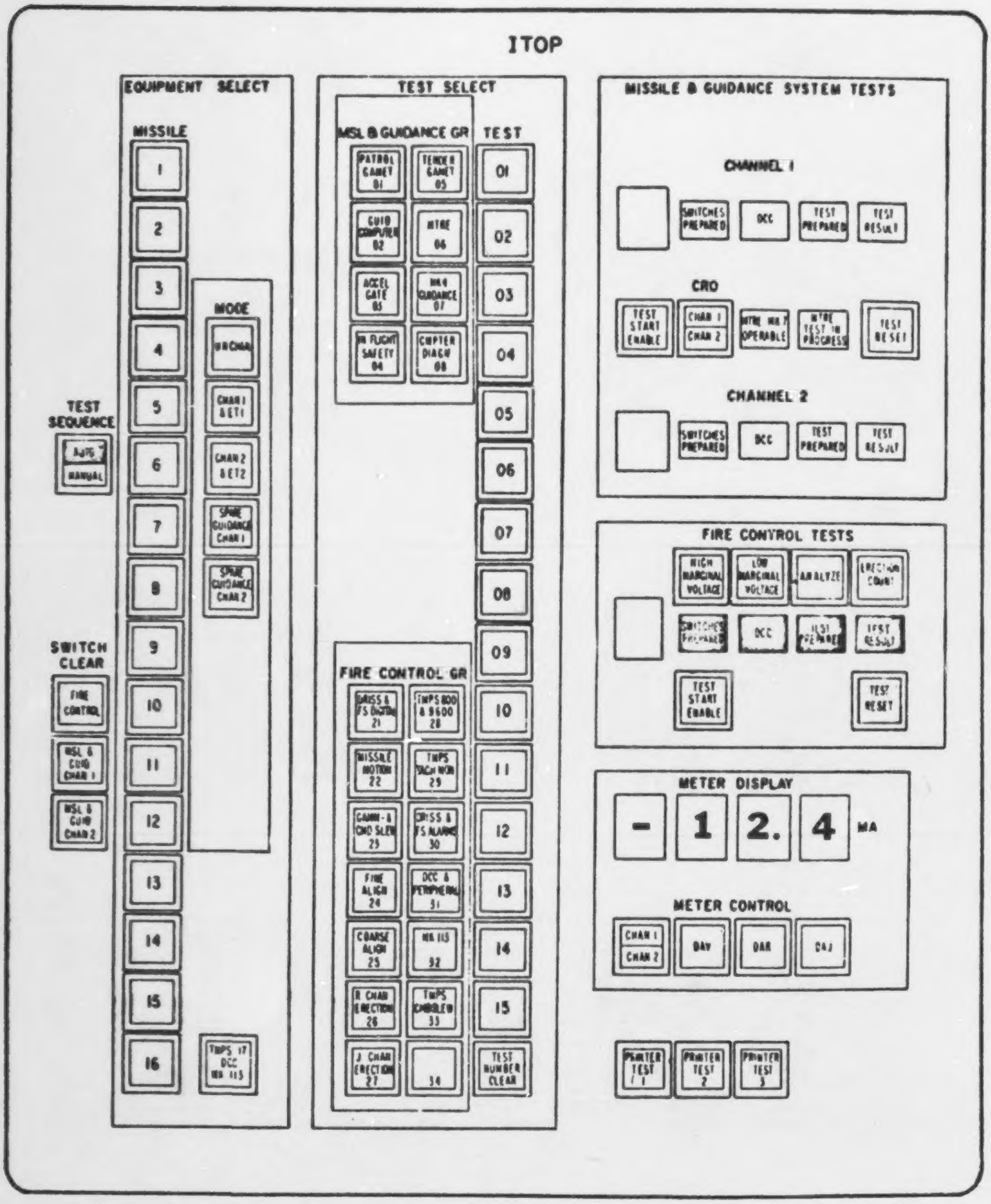
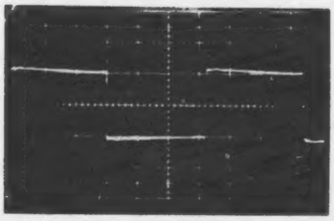
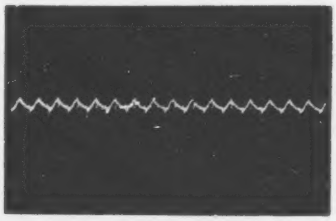

ITOP
EQUIPMENT SELECT
MISSILE
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
MODE
UNCHAN
CHAN 1 BET 1
CHAN 2 BET 2
SPARE GUIDANCE CHAN 1
SPARE GUIDANCE CHAN 2
TMPS 17 DCC MK 113
TEST SEQUENCE
AUTO
MANUAL
SWITCH CLEAR
FIRE CONTROL
MSL & GUID CHAN 1
MSL & GUID CHAN 2
TEST SELECT
MSL & GUIDANCE GR
PATROL GAMET 01
TENDER GAMET 05
GUID COMPUTER 02
MTRE 06
ACCEL GATE 03
MK4 GUIDANCE 07
IN FLIGHT SAFETY 04
CMPTER DIAGN 08
FIRE CONTROL GR
GRISS & FS DIGITAL 21
TMPS 800 & 9600 28
MISSILE MOTION 22
TMPS TACH MON 29
CAMM - & CMD SLEW 23
GRISS & FS ALARMS 30
FINE ALIGN 24
DCC & PERIPHERAL 31
COARSE ALIGN 25
MK 113 32
R CHAN ERECTION 26
TMPS CHBLEW 33
J CHAN ERECTION 27
34
TEST
01
02
03
04
05
06
07
08
09
10
11
12
13
14
15
TEST NUMBER CLEAR
MISSILE & GUIDANCE SYSTEM TESTS
CHANNEL 1
SWITCHES PREPARED
DCC
TEST PREPARED
TEST RESULT
CRO
TEST START ENABLE
CHAN 1
CHAN 2
MTRE MK 7 OPERABLE
MTRE TEST IN PROGRESS
TEST RESET
CHANNEL 2
SWITCHES PREPARED
DCC
TEST PREPARED
TEST RESULT
FIRE CONTROL TESTS
HIGH MARGINAL VOLTAGE
LOW MARGINAL VOLTAGE
ANALYZE
ERECTION COUNT
SWITCHES PREPARED
DCC
TEST PREPARED
TEST RESULT
TEST START ENABLE
TEST RESET
METER DISPLAY
- 1 2. 4
MA
METER CONTROL
CHAN 1
CHAN 2
DAY
DAR
DAJ
PRINTER TEST 1
PRINTER TEST 2
PRINTER TEST 3

Figure 3
Example of Wave Form Page from Data Book

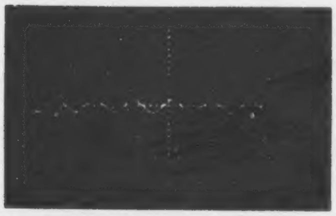
DMS F51 DATA BOOK



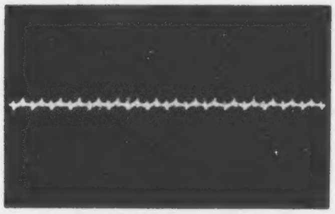
W117 Vert. Sens. - 0. 2 v/cm
Sweep Time - 0. 2 ms/cm



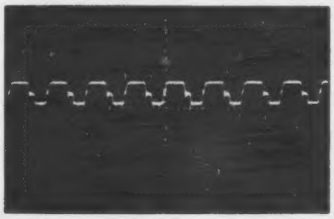
W121 Vert. Sens. - 0. 2 v/cm
Sweep Time - 2 ms/cm



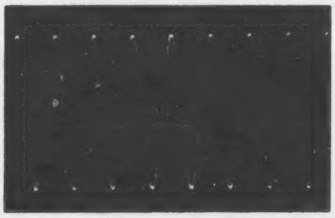
W118 Vert. Sens. - 0. 1 v/cm
Sweep Time - 2 ms/cm



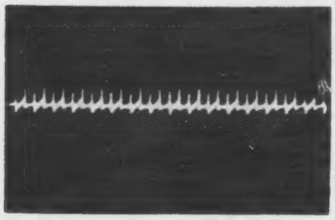
W122 Vert. Sens. - 0. 2 v/cm
Sweep Time - 2 ms/cm



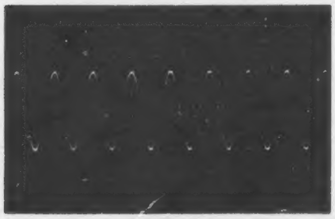
W119 Vert. Sens. - 1 v/cm
Sweep Time - 1 ms/cm



W123 Vert. Sens. - 0. 1 v/cm
Sweep Time - 2 ms/cm



W120 Vert. Sens. - 0. 2 v/cm
Sweep Time - 2 ms/cm



W124 Vert. Sens. - 0. 1 v/cm
Sweep Time - 2 ms/cm

Figure 4
Example of Printout Page from Data Book

DMS F51 DATA BOOK

```
1 2408 CHN     2 FAIL
1       FAIL SIG  001
1           FAILP0001
1           FAILP0001
1           FAILP0000
1           FAILP0001
1           FAILP0001
1           FAILP0002
1           FAILP0001
1           PASSN0004
```

T103

```
2116


5101

2409

2309

2215
2214
2213
2212

2115
2114
2113
2112
```

T102

```
109
```

T101

```
           P

01 2   2409

 +12.4MA F

01 2   2408

 -43.2MA P

01 2   2407

 -44.2MA P

01 2   2406

2206

 +47.0MA P

01 2   2405


 +114.MA P

01 2   2404

2203


 -0.13MA P

01 2   2403

2203


 -0.45MA P

01 2   2402

2203

 -0.03MA P

01 2   2401
```

T104

```
           P

02 2   2409

 -0.04MA F

02 2   2408

 -43.2MA P

02 2   2407

 -3 .3MA P

02 2   2406

2206

 +47.0MA P

02 2   2405


 +114.MA P

02 2   2404

2203


 +0.00MA P

02 2   2403

2203


 -0.42MA P

02 2   2402

2203

 -0.04MA P

02 2   2401
```

T105

appropriate page in the Data Book (which is tabbed for ease in locating the referenced data) and to study the panel display for appropriate information (which is in color in the actual data book).

The Exercise Sheet, shown in Figure 5, controls the branching between items by the use of latent image numbers printed adjacent to each response alternative. As the examinee marks the rectangular area next to the alternative with a latent image developing pen, a number appears in the area. If this number is N000, it directs the examinee to an item (the first item in the Note Book) which tells him/her that the wrong answer was selected and that he/she should go back to the item and select another alternative.[1] When the examinee selects the correct alternative, the number directs him/her to the next item in the Note Book. The items in the Note Book are scrambled so that mere perusal will not enable the examinee to detect the proper sequence of items.

In addition to the Note Book, Data Book, and Exercise Sheet, the examinee has access to the full range of documentation that would normally be available were he/she performing the operational or maintenance procedure on board the submarine.

A computerized version of the DMS has been developed using an interactive terminal to present the items normally found in the Note Book. The Data Book, on the other hand, has been left in its present form. The computer controls the branching between items and provides feedback to the examinees. The Central Test Site has obtained a General Electric Training System (GETS), an interactive terminal with random access slide projection capability. This will make possible full automation of the DMS, by presenting the panel displays, waveforms, and diagnostic printouts of the Data Book by means of either slides or the GETS plasma display. A comparison between this automated version of the DMS and the paper-and-pencil version will be conducted in the near future.

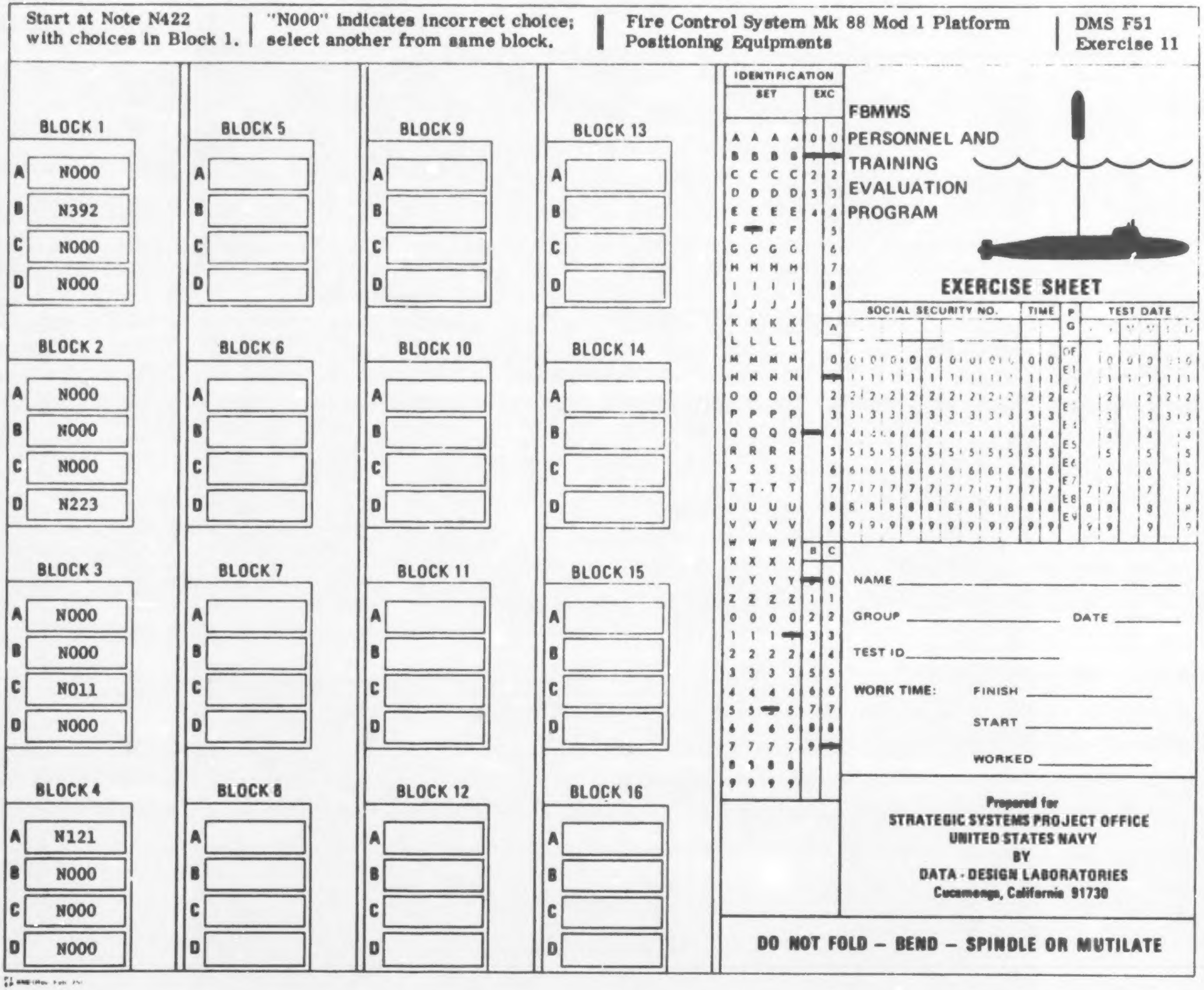
Scoring

The first scoring system developed for the DMS (Braun & Tindal, 1974) was logarithmic in nature, assigning 9, 8, 6, and 0 points for 1, 2, 3, and 4 answers marked for each block, respectively. As scores for Navy technicians were accumulated, both instructors and supervisory personnel commented on the lack of discrimination of the scores obtained. In particular, it was pointed out that there should be a large difference in score points between one and two responses within a block and a successively lesser difference between two and three attempts, and three and four attempts; in other words, an exponential scoring system was proposed. Furthermore, instructors suggested that the proportion of correct initial responses (PCIR) should constitute a major factor in the scoring scheme and that this factor should be applied to further distribute the scores.

---

[1] The Exercise Sheet also explains that N000 is an incorrect choice; note that the first four blocks have been filled in to show the appearance of latent image codes.

Figure 5
Example of Exercise Sheet



Start at Note N422 with choices in Block 1. | **"N000" indicates incorrect choice; select another from same block.** | **Fire Control System Mk 88 Mod 1 Platform Positioning Equipments** | **DMS F51 Exercise 11**

	BLOCK 1	BLOCK 2	BLOCK 3	BLOCK 4	BLOCK 5	BLOCK 6	BLOCK 7	BLOCK 8	BLOCK 9	BLOCK 10	BLOCK 11	BLOCK 12	BLOCK 13	BLOCK 14	BLOCK 15	BLOCK 16
A	N000	N000	N000	N121												
B	N392	N000	N000	N000												
C	N000	N000	N011	N000												
D	N000	N223	N000	N000												

IDENTIFICATION
SET EXC

FBMWS
PERSONNEL AND
TRAINING
EVALUATION
PROGRAM

EXERCISE SHEET

SOCIAL SECURITY NO. | TIME | P G | TEST DATE

NAME ____________________

GROUP ____________ DATE ________

TEST ID ____________

WORK TIME: FINISH ____________

START ____________

WORKED ____________

Prepared for
STRATEGIC SYSTEMS PROJECT OFFICE
UNITED STATES NAVY
BY
DATA - DESIGN LABORATORIES
Cucamonga, California 91730

DO NOT FOLD – BEND – SPINDLE OR MUTILATE

The current DMS scoring system (Hamm & Bradshaw, 1976) is based upon the following propositions:

1. The DMS scores should range between 0 and 100.
2. As the PCIR increases, score points should be earned at an increasing rate.
3. For any item second and third attempt correct responses should not add more score points than can be obtained by selecting the correct answer on the first attempt (a factor of .9 was chosen to satisfy this inequality).
4. For any item the score points earned by the third attempt correct response should not exceed one-half of the maximum score points given for second attempt correct response.

Using these propositions, the following formulas were derived for scoring the DMS:

$$S_t = S_1 + S_2 + S_3, \qquad [1]$$

where S_t = total exercise score;

S_1 = score points earned for items answered correctly on first attempt;

S_2 = score points earned for items answered correctly on second attempt; and

S_3 = score points earned for items answered correctly on third attempt.

The contributions of the three components of the total score are as follows:

$$S_1 = 100\,(e^{.693P} - 1) \qquad [2]$$

$$S_2 = \left(\frac{R_2}{N-R_1}\right)(.9)\left[100\,(e^{.693(P+\frac{1}{N})} -1) - 100(e^{.693P} -1)\right] \qquad [3]$$

$$S_3 = \left(\frac{R_3}{N-R_1}\right)\left(\frac{.9}{2}\right)\left[100\,(e^{.693(P+\frac{1}{N})} -1) - 100(e^{.693P} -1)\right] \qquad [4]$$

where N = number of items in the DMS exercise;

R_1 = number of items answered correctly on first attempt;

R_2 = number of items answered correctly on second attempt;

R_3 = number of items answered correctly on third attempt; and

P = proportion of items answered correctly on first attempt,

i.e., $\text{PCIR} = \frac{R_1}{N}$

A consolidated scoring formula is given as follows:

$$S_{\text{total}} = 100 \left\} e^{\frac{.693R}{N}} -1 \left[1+\left(\frac{.9R_2+.45R_3}{N-R_1}\right)\left(e^{\frac{.693}{N}} -1\right)\right] -1 \right\{ . \qquad [5]$$

Table 1
Number of Attempts for Each of 12 Items for a Sample of 10 Examinees

	Item												Frequency of No. of Attempts			
Examinee	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4
1	1	1	1	1	1	1	1	1	2	2	4	1	9	2	0	1
2	1	1	1	3	1	1	1	4	1	3	4	1	8	0	2	2
3	2	1	1	1	1	1	1	3	1	1	1	1	10	1	1	0
4	1	1	1	3	1	1	2	3	3	2	3	4	5	1	1	0
5	1	1	1	1	1	1	1	4	1	2	1	2	9	2	0	1
6	1	2	1	2	1	1	1	1	1	1	3	1	9	2	1	0
7	1	1	1	1	1	1	1	1	2	2	1	1	10	2	0	0
8	1	1	1	2	1	1	1	1	1	2	1	1	10	2	0	0
9	1	2	2	2	2	4	4	4	2	3	1	1	3	5	1	3
10	1	1	1	2	1	1	1	3	1	3	2	2	7	3	2	0

Table 2
Transformation of Number of Attempts into Score Equivalents

	Item												
Examinee	1	2	3	4	5	6	7	8	9	10	11	12	DMS Score
1	7.57	7.57	7.57	7.57	7.57	7.57	7.57	7.57	3.00	3.00	0	7.57	74.13
2	7.34	7.34	7.34	1.06	7.34	7.34	7.34	0	7.34	1.06	0	7.34	60.84
3	4.77	7.82	7.82	7.82	7.82	7.82	7.82	2.38	7.82	7.82	7.82	7.82	85.35
4	6.70	6.70	6.70	.51	6.70	6.70	1.02	.51	.51	1.02	.51	0	37.58
5	7.57	7.57	7.57	7.57	7.57	7.57	7.57	0	7.57	3.00	7.57	3.00	74.13
6	7.57	3.00	7.57	3.00	7.57	7.57	7.57	7.57	7.57	7.57	1.50	7.57	75.63
7	7.82	7.82	7.82	7.82	7.82	7.82	7.82	7.82	4.77	4.77	7.82	7.82	87.74
8	7.82	7.82	7.82	4.77	7.82	7.82	7.82	7.82	7.82	4.77	7.82	7.82	87.74
9	6.31	.71	.71	.71	.71	0	0	0	.71	.35	6.31	6.31	22.83
10	7.12	7.12	7.12	1.60	7.12	7.12	7.12	.80	7.12	.80	1.60	1.60	56.24
Total	70.59	63.47	68.04	42.43	68.04	67.33	61.65	34.47	54.23	34.16	40.95	56.85	662.21

From Tables 1 and 2, it can be seen that as the PCIR increases, the score equivalents (S_1') increase.[2] For example, Examinees 7 and 8 each had ten correct responses on the first attempt and two correct responses on the second attempt. The first attempt equivalent score is 7.82, and the second attempt equivalent score is 4.77. However, for Examinee 1 (who had nine first attempt correct responses, two second attempt correct responses, and one fourth attempt correct response), the score equivalents are 7.57 for each first attempt and 3.00 for each second attempt. No credit is earned for the fourth attempt within a given block.

Characteristics of Test Items

The items that are used in the construction of the DMS constitute, for the most part, sequences of corrective maintenance procedures. It has been found that a significant number of procedures can be represented in the range of 7-16 items. When a sequence consists of less than seven items, collateral knowledge items are inserted. In some cases knowledge items are inserted, even though the seven skill steps have been attained, in order to give better continuity to the sequence.

Thus DMS items are classified as exemplifying either skill or knowledge. For each category there are two "levels," C1 and C2 for skill and T2 and T3 for knowledge, where T2 knowledge supports C1 skill and T3 knowledge supports C2 skill. A complete list of the categories of skill and knowledge as used in the FBM training program is given in Appendix A.

There is an obvious analogy between the Navy's designations (T2, T3, C1, C2) and some of the current taxonomies of educational objectives (e.g., Bloom, Hastings, & Madaus, 1971). The T2 and T3 items can be regarded as "knowledge" and "understanding " respectively, whereas C1 and C2 items can be regarded as the "application of knowledge" and the "application of understanding" respectively.

Validity of the DMS

Criterion-Related Validity

The equipment on the FBM submarines is extremely reliable. When an infrequent failure occurs, the most capable technician is assigned to direct a team to effect the repair. It is therefore difficult, if not impossible, to obtain criterion performance measures for the individual technician.

A true performance measure of corrective maintenance most corresponds to the "prefaulted module" (PFM) test, which has received widespread acceptance in the FBM community. In this test, one of the modules of a piece of electronic equipment is faulted in a manner not discernable by observation. The technician is required to isolate the fault to a replaceable module (circuit board or plug-in unit). While performing the trouble-shooting procedure, he/she is observed by an instructor who records the steps, remedies (module swaps), and time required to locate the fault. The examinee is given a score based upon a set of weights assigned to these three parameters (Hamm & Bradshaw, 1975).

---

[2] The score equivalents are $S_1' = \frac{S1}{R1}; \ S_2' = \frac{S2}{R2}; \ S_3' = \frac{S3}{R3}$

In a recent study of the Analog/Digital Converter Maintenance Course (for Central Navigation Computer trainees), the results obtained from PFM and DMS exercises covering the same areas (but not the identical faults) were correlated. A Pearson r of .524, which is significant at the .01 level for the number of examinees (N=28), was obtained.

Concurrent Validity

Evidence of concurrent validity is cited in a study by Braun, Robinson, and Tindall (1975). Missile technicians who had hands-on experience during a patrol were compared to missile technicians who did not have hands-on experience with a particular piece of test equipment (MTRE-7). A post-patrol DMS, which dealt with the use of the MTRE-7 in diagnosing equipment faults, revealed a significant difference in the predicted direction for technicians having direct experience with the MTRE-7 (p<.01 for a one-tailed test).

An examination of the relationships between DMS scores and paygrades revealed a characteristic curve. There was a near linear trend for the first three paygrade categories (E3/E4, E5, and E6) and a leveling of the curve for E7/E8. An example of such a curve is found in the data for a DMS derived from a recent System Achievement Test (F17603), as shown in Table 3.

Table 3
Mean and Standard Deviation of DMS
Scores by Paygrade

Paygrade	N	Mean	$S.D.$
E3/E4	69	166.93	48.26
E5	122	191.25	39.57
E6	45	215.71	38.64
E7/E8	31	219.90	36.96

The differences between the means for E3/E4 and E5 was 24.32, which was significant at the .01 level (t=3.76 for 189 df). The difference between E5 and E6 was 24.46, which was also significant at the .01 level (t=3.56 for 1[illegible] df). However, the difference between E6 and E7/E8 was only 4.19, which was not significant (t=.47 for 74 df). The reason for the leveling of scores in the higher paygrades is the fact that a large proportion of E7/E8 technicians are charged with administrative tasks, while the E5/E6 technicians have direct responsibility for the equipment.

Content Validity

As noted above, there are two general types of items in the DMS: skill (C1 and C2) and knowledge (T2 and T3). Since the DMS is constructed as a skill-related test, it should be expected that there would be a significant difference in proportion of correct responses for C1 and C2 items and a lesser difference in proportion of correct responses for T2 and T3 items.

Table 4 gives the mean number of answer attempts per block for C1 versus C2 items for a selected DMS. The average for C1 items was 1.32 attempts per item, while the average for C2 items was 1.80 attempts per item. The difference was found to be significant at the .002 level for a directional test.

Table 4
Mean and Standard Deviation of
Number of Attempts for Skill Items
(*N*=243)

Skill Level	No. of Items	Mean	*S.D.*
C1	9	1.32	.25
C2	14	1.80	.45

Note. Mean difference of .48 was statistically significant at the .002 level for a one-tailed test.

Table 5 gives the mean number of answer attempts for T2 and T3 items of the same DMS. The average for T2 items was 1.45 attempts per item, while the average for T3 items was 1.60 attempts per item. The difference was found to be nonsignificant.

Table 5
Mean and Standard Deviation of
Number of Attempts for Knowledge Items
(*N*=243)

Knowledge Level	No. of Items	Mean	*S.D.*
T2	6	1.45	.33
T3	9	1.60	.34

Note. Mean difference (.15) not statistically significant at $p \leq .05$.

Table 6 gives the mean time per item for all DMS exercises combined. A C1 exercise has a preponderance of C1/T2 items, while a C2 exercise has a preponderance of C2/T3 items. Data collected over a period of several years and based upon more than 4,000 examinees yielded an average time per item for C1 exercises of 2.07 minutes and an average time per item for C2 exercises of 2.92 minutes. The difference was found to be significant at the .001 level for a nondirectional test. These data provide additional evidence of content validity of the DMS.

Table 6
Mean Time Per Item for DMS Exercises
($N \geq 4000$)

Exercise Level	No. of Exercises	Mean Time	*S.D.*
C1	29	2.07	.79
C2	84	2.92	.77

Note. Mean difference of .85 was statistically significant at the .001 level.

Reliability of the DMS

No large-scale studies of the reliability of the DMS have been conducted as yet. The determination of test-retest reliability is precluded by a Navy requirement to use a different set of DMS exercises for successive tests. Thus, it was necessary to confront the problem of estimating its internal consistency reliability.

As noted in the section on scoring the DMS, a score matrix, representing the score equivalents of the number of attempts per item,can be generated. These score equivalents (Table 2), which add up to the scores given to each examinee, can be used to estimate internal consistency reliability of the DMS. It is assumed that the items of the DMS can be treated independently, since all the necessary antecedent information is presented with each item at the time it is encountered by the examinee. Using several small random samples from the data obtained from a recent DMS, internal consistency estimates ranging from .75 to .88 were obtained.

Elaboration of the DMS

At the present time, a uniform format is used for the DMS, i.e., a "single-thread" model with up to 16 items. This means that each item brings the examinee back to the "right" track without allowing him/her to veer off course. A possible elaboration of the DMS is a multiple-thread model which could provide greater fidelity of simulation, but complicate the production and scoring.

There are two reasons why there has been no departure from the present format. The first is related to the manner in which the DMS is scored, i.e., the exponential scoring scheme. In this scheme the difference between first attempt correct response and second attempt correct response is greater than that between second attempt correct response and third attempt correct response. Thus, if branching occurred from one item to a second item, increasing the number of alternatives from four to eight, the differential scores would become successively smaller.

A second reason for persisting with the present format lies in the logistics of test administration. A one-page answer sheet has been designed which is extremely legible and relatively free from clerical errors. In addition, although the proctor reviews each form, the answer sheet can be machine scored by having the examinees fill in the mark-sense slots next to the items and using optical or mark-sense readers. If a more complex model were to be used, either an answer sheet with a greater density of blocks or two answer sheets would be needed. In either case, this would complicate the logistics of test administration and scoring.

Conclusion

The DMS, as presently constituted, has received a high degree of examinee acceptance and gives evidence of reasonable levels of validity. The DMS in its present form is not an "adaptive" test,but rather an "interactive" test. The DMS provides immediate knowledge of results, which (as Betz and Weiss, 1976, point out) generally makes a test more interesting and stimulates better performance. In its present form, the DMS can be integrated into a self-study course: There is initial testing in the form of one set of exercises, further study, and follow-up testing with another set of exercises from the same DMS.

With the increased availability of computers at the various Navy training sites, it may be possible to utilize a computerized version of the DMS at the land-based facilities while retaining the paper-and-pencil version for use on ships and submarines. Furthermore, with the phenomenal growth of microprocessors, computerized training and testing devices will likely almost completely supplant the use of written materials in the near future.

References

Betz, N. E., & Weiss, D. J. Psychological effects of immediate knowledge of results and adaptive ability testing (Research Report 76-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1976.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. Handbook on formative and summative evaluation of student learning. New York: McGraw-Hill, 1971.

Braun, F. B., & Tindall, J. E. A new sequential multiple choice skill related testing device. Proceedings of the 16th Annual Conference of the Military Testing Association, Oklahoma City, OK, 1974.

Braun, F. B., Robinson, M. A., & Tindall, J. E. The decision measurement system--current status. Proceedings of the 17th Annual Conference of the Military Testing Association, Indianapolis, IN, 1975.

Hamm, E. L., & Bradshaw, C. W. Development of performance test scoring procedures (PTEP Special Study Report 7-75). Norfolk, VA: Data-Design Laboratories, 1975.

Hamm, E. L., & Bradshaw, C. W. Review of Decision Measurement System scoring methodology (PTEP Special Study Report 1-76). Norfolk, VA: Data-Design Laboratories, 1976.

APPENDIX A

NAVORD OD 43180 REVISION 2 (VOLUME 6, PART 1)

Training Objective Statements
for FBM Weapon System Technicians

T0 Statement T0 includes the background skill and knowledge which is prerequisite to the understanding of the operation and maintenance of the system/subsystem/equipment. This level of training includes:

B1 Completion of training provides the level of knowledge necessary to recognize or recall ideas, phenomena, symbology, and terminology which are prerequisite to the comprehension of a task or function.

B2 Completion of training provides the comprehension of the principles, rules, and concepts necessary to solve given problems and situations of an assigned task or function.

S Completion of training provides the ability and knowledge to apply principles, rules, or concepts in the solution of given problems and performance of assigned tasks or functions.

B - Basic Knowledge
S - Basic Skill

Training Objective Statements for FBM Weapon System Technicians (Continued)

F1	Completion of training provides familiarity with the purpose, function, and location of a specific system/subsystem/equipment, on a subsystem/equipment level, and with supporting documentation, required to safely perform general duties within the FBM Weapon System. When applied to Computer Software the following definitions shall apply: Completion of training provides familiarity with documentation, purpose and function of the software package on a level required to safely perform general duties within the FBM Weapon System.
T1	Completion of training provides the depth of knowledge to understand functional operation and to support performance of all normal operational tasks at the O1 level. This knowledge also supports the T2 level.
T2	Completion of training provides the depth of knowledge to understand functional operation and to support performance of all casualty and normal operational tasks at the O2 level, all preventive maintenance at the P1 level, and corrective maintenance at the C1 level. The knowledge also supports the T3 level.
T3	Completion of training provides the depth of knowledge to support all corrective maintenance at the C2 level.
O1	Completion of training provides the skill and knowledge to perform, with supervision, normal operational procedures. This skill level supports the O2 level.
O2	Completion of training provides the skill and knowledge to perform, with supervision, casualty operational procedures and those normal operational procedures requiring advanced analysis. The ability to perform O1 level procedures, without supervision, is gained through experience.
O3	This training level provides the skill and knowledge to perform, without supervision, all casualty and all normal operational procedures. This skill level is gained through experience.

Training Objective Statements for FBM Weapon System Technicians (Continued)

P1	Completion of training provides the skill and knowledge to perform, with supervision, preventive maintenance procedures.
P2	This training level provides the skill and knowledge to perform, without supervision, preventive maintenance procedures. This skill level is gained through experience.
C1	Completion of training provides the skill and knowledge to perform, with supervision and to the authorized maintenance level, documented fault isolation and repair procedures. This skill level supports the C2 level.
C2	Completion of training provides the skill and knowledge to perform, with supervision and to the authorized maintenance level, repairs and isolation of faults that cannot be located using procedures contained in prescribed documentation. Some training on maintenance covered by documented procedures may be classified as C2 when advanced analysis is required to complete the procedure. The ability to perform C1 maintenance procedures, without supervision, is gained through experience.
C3	This training level provides the skill and knowledge to perform, without supervision, diagnosis of equipment malfunctions, fault isolation, and all repairs. This skill level is gained through experience.

DISCUSSION: SESSION 4

ERNST Z. ROTHKOPF
BELL LABORATORIES

The papers by McGuire, Knerr, and Robinson and Walker dealt with adaptive tests and the application of adaptive testing to the measurement of competence. The problems that were discussed in these papers were of great interest to me because I have been working in this area for a long time. In the mid-1950s I was doing research at the Air Force Personnel and Training Research Center, which developed a simulation known as the TAB test. The TAB Test, developed by Glaser, Damrin, and Gardner (1954), was concerned with the simulation of problems with a superheterodyne radio set.

The TAB Test provided testing options to troubleshooting technicians. The technician chose one of these options, removed a tab, and was shown the consequences of the action taken (e.g., a voltage reading, an oscilloscope display). The TAB Test had an ingeniously simple sequence-retrieval system built into it. After the testee removed the tab, the tab was placed on a spindle. In that way, it was possible to recover afterward not only those particular troubleshooting choices that had been made but also the sequence in which these actions occurred. Subsequently, Crowder (1959) developed intrinsic programmed instruction from the TAB Test, including scrambled books initially developed to simulate troubles in various bombardment-navigational systems and, later, to simulate bridge-playing. The simple spindle tab test provided information at least as interesting and complex as some of the computer systems which have been discussed at this conference.

Much of the basis of the technology that was described in these three papers existed in 1956. What has changed in the interval? What advances have been made and what has been accomplished?

A number of things did happen, as demonstrated very clearly by Christine McGuire's paper. First, a very elaborate technology for simulation has been developed. McGuire has taken advantage of this technology to create tests that simulate more realistically and to provide situations that come very close to meeting the criteria she has in mind. Second, the computer has developed during this interval; but I cannot say--on the evidence of these three papers--that the computer has been used in a way commensurate with its powerful capabilities.

We seem to be in the same stage of debate about simulation and the adaptive assessment of competence as we were a few years ago when people kept saying that computer-aided instruction was used simply as a high-class page turner. It seems as though this same profound level of discussion has now been reached in the adaptive testing area. The computer, as far as I can see, is not often used in an interesting and ingenious manner to carry out adaptive achievement testing. This is not a criticism of any of the present papers;

they were, as a matter of fact, for the most part very sensible. Not much use was made of computers in these studies. The investigations used whatever was needed, which, in many cases, involved only paper simulation and various other simulation techniques.

There is a problem here, however: Something has not happened. No really interesting ideas have been presented in this particular area about how the computer can be applied in a useful, practical, different, and more powerful way to the problems related to the measurement of competence. How to use the computer most effectively for this type of measurement is a problem that has not yet been solved and may not even be solvable in the immediate future.

There is another problem that is inherent in all of the papers; it has to do with the question of purpose. McGuire's paper definitely expressed purpose: She was concerned with the certification of health professionals. The test data obtained by her can be used by state certifying authorities in some way. It has obvious face validity. The relevance of her simulations is probably even defensible from a legal point of view.

The purposes of Knerr, and Robinson and Walker were not as clear. There are two possible interpretations: Either they wanted to provide some kind of training experience for people or, conceivably, to evaluate individuals prior to their going on a particular kind of duty, perhaps in order to decide on needed remedial training. If what they were concerned with involved remediation, both of those tests were rather poorly constructed, since the tests did not yield any information sufficient for this purpose.

Robinson and Walker seemed to be saying that the sequence of testing an individual has to go through is determined by the logic of the hardware system. This is an ideal condition for the construction of certain simulation devices. It is also an ideal condition for the elimination of a job. If a forecast can be made of what a particular testing sequence ought to be and how to evalute the results in order to locate the trouble, then a machine can probably be produced to locate the fault. This suggests that the proposed new testing methodology would work best under conditions that are also conducive to the elimination of the job for which training and testing procedures are being developed. I would think that one would be very pessimistic about entering that kind of work.

Knerr's paper raised a different problem about purpose. It is whether or not a utility function of the kind he proposed can be stated without specific ideas about (1) what the trainee has to do after the completion of training and (2) how a particular equipment system is going to be used. Ideas about utility involve such matters as cost, the amount of work to be done, the amount of replacement value involved, and risks about wrong decisions. A closed system of that particular kind requires some fairly strong assumptions about how the individual is going to do the job and about the ability of the so-called expert to generate the proper way of handling the system. If the expert has considerable knowledge, it is likely that the system has been in existence for some time. However, training and diagnosis usually are most needed (particularly in military systems) when a system is quite new and many people must be trained to operate it.

It is with such new systems that there is the greatest difficulty in providing the expert model for the student. With new systems, there is no history of troubleshooting and no knowledge about utilities. Any utility or any choice point in a troubleshooting sequence ought to be weighted not only with the number of possibilities but also with the likelihood that a particular fault will be located in a particular place in a working system.

Nevertheless, there are some interesting ideas here. Even if we cannot conceive of a better task for the computer than page turning, at least the computer can keep a record and retain it for a long time. This record could be analyzed retrospectively. That is, the computer can add to our knowledge of how a skill is learned by providing a record of people trying to exercise that skill. This is a unique feature of the computer. It probably ought to be exploited a lot more fully than it is at present. Adaptive features probably have their most interesting future in the simulation setting--exploring the difficulties the testee may be having in serving as a diagnostic tool. A global test problem in a standard achievement test setting provides few opportunities to explore deficiencies in knowledge or skill. Adaptive achievement tests, on the other hand, make the most of the information that the testee supplies. An adaptive diagnostic screening may provide very specific recommendations for remediation without undue demands on recommendations for the student's time.

Record keeping and the specific diagnosis of needed remediation present strong possibilities, augmenting some of the testing techniques that have been described here. There is also clearly a need for better psychological understanding of subject matter. What conclusion can we draw from the fact that after more than 20 years, so very little has happened to change our understanding of how people perform difficult problem-solving tasks? Why is there so little information coming from the human factors side (the psychological side)? Why is it that what is held seductively before contract offices is really the smell of hardware, bent aluminum, terminals, and futuristic visions of 21st century control stations? The problem is, in part, that researchers avoid doing a certain kind of needed study. They like doing methods studies, regardless of how fruitless they are. They love to play with the statistics of test construction and with new techniques for presenting information to people. But they do not really like to embark on the detailed task analysis of common skilled human tasks. Although task analyses are really needed, there is little glory in doing them; and such work tends to be difficult to publish. There ought to be room for it in professional journals and in the journals of industrial and military groups. There is a need for the detailed analysis of technological subject matter--how people perform difficult jobs, how they troubleshoot, how they manage their work. This may seem prosaic, but it is potentially powerful information and would be extremely valuable to anyone embarking on a program of adaptive testing of competence.

References

Crowder, N. A. Automatic tutoring by means of intrinsic programming. In E. H. Galanter (Ed.), Automatic teaching: The state of the art. New York, NY: John Wiley & Sons, 1959, 109-116.

Glaser, R., Damrin, D. E., & Gardner, F. M. The Tab item: A technique for measurement of proficiency in diagnostic problem solving tasks. *Educational and Psychological Measurement*, 1954, *14*, 283-293.

Session 5

Implementations of Adaptive Testing

Computerized Adaptive Testing and Personnel Accessioning System Design	Mark A. Underwood Navy Personnel Research and Development Center
Implementation of a Model Adaptive Testing System at an Armed Forces Entrance and Examination Station	Malcolm J. Ree Air Force Human Resources Laboratory
Computerized Adaptive Testing with a Military Population	Steven Gorman Headquarters Marine Corps
Implementation of Tailored Testing at the Civil Service Commission	Richard H. McKillip U.S. Civil Service Commission
Operational Considerations in Implementing Tailored Testing	Harold Segal U.S. Civil Service Commission
A Low Cost Terminal Usable for Computerized Adaptive Testing	J. P. Lamos and B. K. Waters Air Force Human Resources Laboratory

SESSION 5: ABSTRACTS

COMPUTERIZED ADAPTIVE TESTING AND PERSONNEL ACCESSIONING SYSTEM DESIGN

MARK A. UNDERWOOD

This paper presents the rationale for establishing the components, performance specifications, software and hardware characteristics, and cost considerations of providing an adequate computing capability for adaptive testing within the context of the military personnel acquisitions environment. Several computer programs that have been developed for adaptive testing are analyzed, with the aim of drawing implications for the system's architecture requirements. A prototype laboratory configuration is described. The special hardware and software requirements of graphics presentation upon ordinary video monitors are identified, and a technique is proposed which utilizes existing military air-to-ground communications technology. Research plans for computer scientists and engineers responsible for prototype development are outlined, and indices of testing cost-effectiveness are estimated for the production version of the prototype system.

IMPLEMENTATION OF A MODEL ADAPTIVE TESTING SYSTEM AT AN ARMED FORCES ENTRANCE AND EXAMINATION STATION

MALCOLM J. REE

A model adaptive testing system was developed and installed in the Armed Forces Entrance and Examination Station (AFEES) in San Antonio, Texas, in order to study its feasibility for use in a military setting. Technical problems confronted included choice of (1) subjects, (2) item selection algorithm, (3) an item-scoring ability estimation algorithm, (4) a programming language, (5) item display techniques, and (6) transmission lines. Administrative problems included (1) training of personnel and (2) simplification of operations, so that examinees do not inadvertently terminate testing. Several solutions to these problems are described. The feasibility of adaptive testing will be investigated by assessing (1) whether or not the system operates efficiently and (2) whether or not adaptive testing is as valid as paper-and-pencil testing.

COMPUTERIZED ADAPTIVE TESTING WITH A MILITARY POPULATION

STEVEN GORMAN

To measure the effectiveness of computerized adaptive testing using the full range of ability, two ability tests will be administered in an adaptive mode via cathode ray tubes with entering Marine recruits as the target population. The testing strategy used in this study will be Owen's Bayesian sequential procedure. Whereas previous studies of Bayesian adaptive ability testing involved college graduates or simulated examinees, this study will include subjects with only 10 years of formal schooling and will cover a broad range of ability. Possible benefits of implementing adaptive testing, and practical problems and their psychometric implications are discussed.

--Continued on next page--

ABSTRACTS (CONTINUED)

IMPLEMENTATION OF TAILORED TESTING AT THE CIVIL SERVICE COMMISSION

RICHARD H. MCKILLIP

The four stages in the development of computerized tailored testing procedures at the Civil Service Commission will be: (1) the PACE test will be administered in the tailored version; (2) the tailored testing capability will be extended to all Civil Service examinations; (3) a comprehensive set of ability constructs will be identified, which will be useful in all employment settings; and (4) banks of items testing these constructs and a system for relating banks to jobs will be developed. The examining system will first tailor item banks to jobs and then tailor tests to individuals. Methods for achieving these objectives, additional possibilities for new applications of the tailored test system, and practical aspects of these applications are described.

OPERATIONAL CONSIDERATIONS IN IMPLEMENTING TAILORED TESTING

HAROLD SEGAL

Implementing tailored testing in a large-scale organizational setting requires a special study to provide the information on which to base key managerial decisions. The U.S. Civil Service Commission is preparing to carry out a feasibility study of this nature in which budgetary and administrative considerations, as well as the technological advantages of the new testing mode, will be analyzed. The results of the feasibility study are intended (1) to supply management with a basis for ascertaining how examining procedures would operate under tailored testing, compared with how they operate now; (2) to provide time frames for successive steps in implementation; and (3) to project resource requirements. Alternative strategies for proceeding with each of these steps will be provided.

A LOW COST TERMINAL USABLE FOR COMPUTERIZED ADAPTIVE TESTING

J. P. LAMOS AND B. K. WATERS

Computerized adaptive testing (CAT) research has grown rapidly since the late 1960s, with extensive investigation into model development, simulated and empirical validation studies, and limited cost analyses. The vast majority of these studies have used large time-sharing computer systems with expensive cathode-ray-tube (CRT) terminals to present test items. Where computer hardware is used for many different functions, as in academic environments, the cost of using these CRTs may be justified; in many envisioned operational implementations of CAT, however, the large system CRT is too expensive, too inflexible, or too immobile to be a feasible testing medium. This paper describes one such operational situation and a developmental effort to design, build, and evaluate an alternative delivery medium.

COMPUTERIZED ADAPTIVE TESTING AND PERSONNEL ACCESSIONING SYSTEM DESIGN

MARK A. UNDERWOOD
NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER

Psychometrics has become increasingly dependent upon computer technology. This is evident in the widespread use of packaged statistical analysis programs, especially the Statistical Package for the Social Sciences (SPSS). Psychologists, in particular, are reared in the environment of the large-scale computer (e.g., IBM 370, CDC Cyber series, Honeywell 6000, Burroughs 6700, PDP-10, UNIVAC 1100/xx), which is amenable to manipulation of large data areas and large programs with appetites for both input/output (I/O) and central processor resources. It would be safe to say that the computer that could adequately satisfy the requirements of the largest statistical programs has yet to be designed. Many of the habits psychometricians have acquired in large computer system usage, however, cannot be tolerated on smaller systems; yet, it is the smaller computer system that will likely provide the apparatus for adaptive testing where it is needed. Therefore, this paper discusses the implementation of adaptive testing and related applications upon small computer systems on a widespread scale.

Background

The Navy Personnel Research and Development Center (NPRDC) has been studying the implications of distributed computing technology for personnel accessioning system design. The availability of lower-cost computer systems makes it possible to configure multi-computer networks which both increase local computing capabilities and reduce network operating cost, especially in the area of telecommunications. Under the mantle of Project CONTRACT (Computerized Navy Techniques for Recruiting, Assignment,Counseling, and Testing), a complex network design has been intensively investigated. This design consists of both fixed and mobile computer systems. These two computer system types are distinguished not only by the requirement of portability, but also by their architectures and capabilities. The mobile system is the one most relevant to the present discussion; it has been configured to perform adaptive testing functions. It is designed also to provide computerized career information, such as NPRDC demonstrated at a high school in San Diego. It is assumed that testing may eventually occur not only at the Armed Forces Entrance Examining Stations (AFEES) but also in the field--at high schools, community colleges, and,on special occasions,shopping centers and exhibitions.

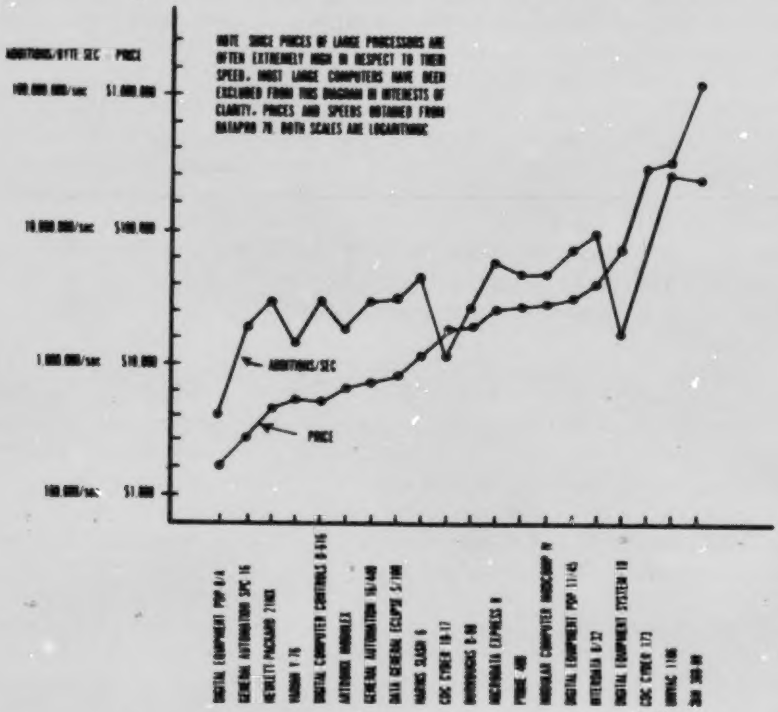
Field testing of this kind requires mobility, because the cost-effectiveness of a computer system physically situated at any one location would be lower than one which could be moved to the location of the next group of people to be tested. Further, the mobile system must be capable of more than adaptive testing. Information dispensing (as well as collecting, in which category testing falls), counseling, quotas, shipping schedules, policies, and perhaps even some criterion-referenced measures are all processes that must be accomodated by the computer system. The mobile system can be interfaced nightly with fixed station computer systems so that local recruiters can follow up on information gathered in the field. Of course, the mobile system need not be mobile; it could be placed in a small office space, such as would be the case at an AFEES.

There are two obvious dimensions for assessment of candidate system configurations: cost and performance. In the course of this discussion, other dimensions will be introduced. However, these two obvious dimensions are important configuration parameters. Cost and performance are intimately related, even with the advent of widespread low-cost microcomputers and their accompanying microperipherals. The breakthroughs have been made in lower memory prices and in the availability of low-cost, low-performance devices which can enable local data collection formatting, editing, and communications. The minicomputer was at the forefront of this trend in the first part of this decade, and the microcomputer will probably obtain the most visibility in the second half. Matching adaptive testing and career information applications to appropriate hardware configurations inevitably calls for compromise between cost-conscious systems analysts and psychometricians; an exceedingly close match between configuration characteristics and applications program resource requirements is essential.

Figure 1 shows the speed vs. entry level system cost of increasingly fast computer systems. There can be no doubt about the possibility of overkill; a program which does not require instantaneous "turnaround" makes use of the same hardware as those which do require it. Programs which are mostly I/O oriented may run no faster on machines with extremely fast central processing units (CPU) because CPU time is not a primary determinant of time to complete execution. The possibility of underkill is obvious; the job is not finished before it is needed. The requirement for multiprogramming, made essential by the multiple-user nature of adaptive testing, introduces the critical factor of the sophistication of the computer system's operating system software. System throughput and resource utilization intelligence decreases dramatically as does both hardware speed and cost.

The level of software sophistication is quite high. For instance, the area of data base structures required to support multiple ability banks, with the possibility of differential starting points based upon ability measured in prior ability dimensions, could become relatively sophisticated. In the worst case, the data base would have to be a fully inverted file, with the key fields the actual raw item parameters as well as item numbers for keys. Furthermore, multiple banks (or files) would have to be interrelatable upon demand.

Figure 1
Selected Entry-Level System Speeds vs. Costs



NOTE: SINCE PRICES OF LARGE PROCESSORS ARE OFTEN EXTREMELY HIGH IN RESPECT TO THEIR SPEED, MOST LARGE COMPUTERS HAVE BEEN EXCLUDED FROM THIS DIAGRAM IN INTERESTS OF CLARITY. PRICES AND SPEEDS OBTAINED FROM DATAPRO 78. BOTH SCALES ARE LOGARITHMIC

On the Use of a Minicomputer

Hardware and software performance monitors are techniques for studying some parameters of program behavior. If a well-written program is analyzed using such tools, quantitative information about computer resource use can be gained. For adaptive test administration software, statistical distributions of machine instruction type can recommend certain architectures. For instance, the occurrence of floating point hardware instructions would indicate the possible value of specialized hardware for floating point--beyond the "floating point boxes" provided by the minicomputer vendors. In a more sophisticated analysis, such studies can reveal patterns of I/O, memory referencing, and class of machine instruction. Hardware and software monitors are an integral part of the effort of matching system functional specifications to hardware and software characteristics.

System requirements. The rationale for the selection of a multi-minicomputer design for a mobile counseling and testing system is worthy of discussion because of the variety of design alternatives available. In addition to the low-cost motivation, these factors led to the identification of a system configuration with the following characteristics:

1. Need for computer control of the graphic items;
2. Capability for performing the adaptive sequencing algorithm locally without deferred scoring or recording;

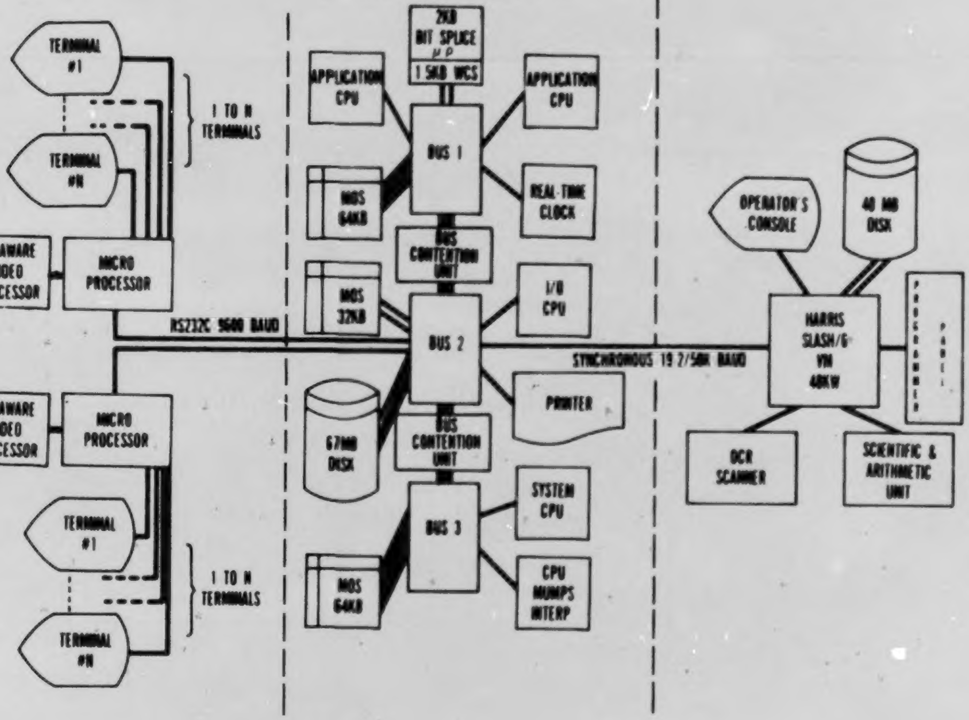
3. Capability for tracking sequences of polychotomously scored items;
4. Increased requirement for test security, allowing for a large item pool to reduce the probability of two individuals with equivalent ability and/or educational backgrounds receiving identical items;
5. Reduce distractors by presenting only one item at a time in the stimulus environment, when necessary;
6. Capability for a large multiple-bank item pool with a complex data structure for rapid retrieval and sophisticated ordering algorithms;
7. Capability for updating the ability estimate after every item presentation, using the most sophisticated strategy (a worst case situation, in terms of computing requirements); and
8. Considerable expansion capability in terms of hardware and software features and also in terms of the number of sophisticated algorithms for stimulus presentation and response data collection and analysis.

Alternative hardware. Utilization of off-the-shelf hardware and software, where performance and cost were in the orders of magnitude needed, simply has not arrived in a sufficiently low-cost package to warrant incorporation into adaptive testing stations. Whether this will change depends upon developments in artificial intelligence and an improved understanding of semantic processes. Even then, it can be certain that this will result in the implementation of highly sophisticated algorithms for language processing which would add significantly to an already burdened small computer system that was chosen for its relative simplicity of design and low cost. The area of video image processing seems to hold the most promise because it presents a means of automatic image digitization and manipulation without excessive expense.

In addition to the high probability of mechanical failure and limited storage capabilities in some cases, microfilm and slide image presentation packages lack the dimension of computer-controlled image manipulation and storage; sensory psychologists would surely certify the importance of control of the image. A look at current item banks shows that photographic-quality images are not frequently used; for instance, a drawing of a vise is used instead of a photograph which would more closely resemble what would be encountered in a military shop. The technology described possesses the capability of presenting "snapshots" from video cameras; a text or discussion could accompany snapshots, or items could consist of sequences of snapshots. The implications for item construction are significant and many. Thus, both hardware and software must be assessed in the identification of a suitable architecture for adaptive test and career information computing.

A prototype. Figure 2 is a prototype configuration under consideration at NPRDC which utilizes multi-vendor, multi-processor technology. It is expected to be able to perform many simultaneous test administrations and, at the same time, is capable of some heavy duty statistical and scientific computing. Also, its use in large quantities would place it in the middle to high end of the mini-computer category. Note that specific functions have been allocated to the various processors in the configuration, as indicated by labels in the processor boxes. Unlike the PDP-11 UNIBUS structure, the Modulex system is capable of mulitiple bus structures as well as multiple processors, and it has twice the rated bus speed. Therefore, it is less likely to become bogged down from movement of data on the bus between processors, from memory to processor, and from peripherals to memory or disk. Aside from the not insignificant task of multi-bus and multi-processor coordination, however, the Artronix operating system is unremarkable, particularly with respect to memory management and higher-level languages.

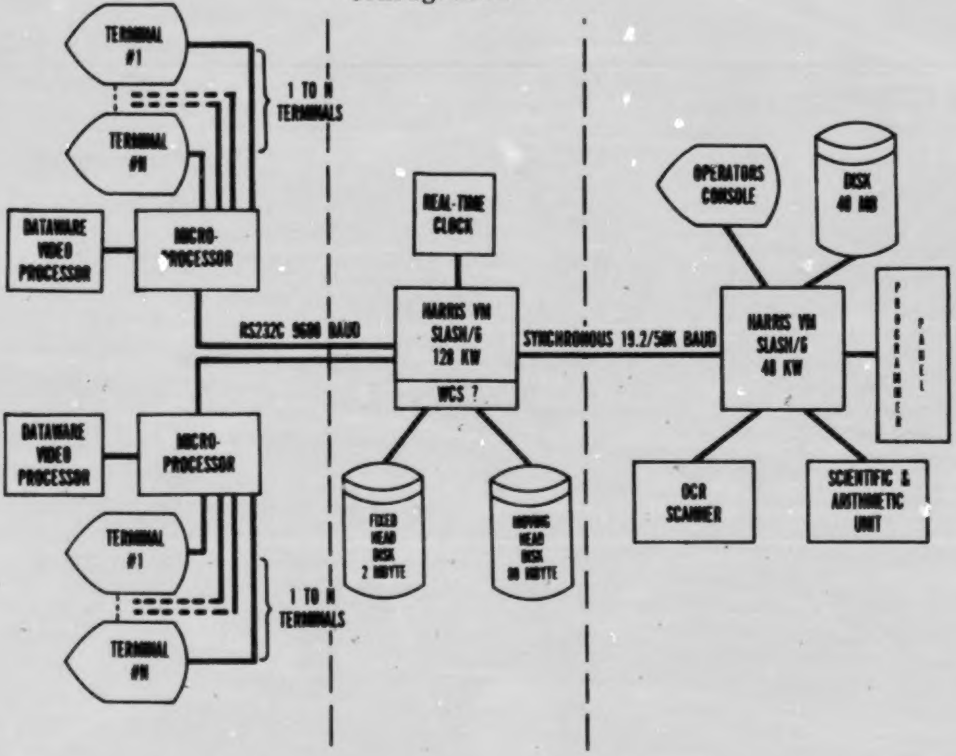
Figure 2
Configuration A


TERMINAL #1
TERMINAL #N
1 TO N TERMINALS
DATAWARE VIDEO PROCESSOR
MICRO PROCESSOR
RS232C 9600 BAUD
DATAWARE VIDEO PROCESSOR
MICRO PROCESSOR
TERMINAL #1
TERMINAL #N
1 TO N TERMINALS
2KB BIT SPLICE μP
1.5KB WCS
APPLICATION CPU
APPLICATION CPU
BUS 1
MOS 64KB
REAL-TIME CLOCK
BUS CONTENTION UNIT
MOS 32KB
I/O CPU
BUS 2
67MB DISK
PRINTER
BUS CONTENTION UNIT
SYSTEM CPU
BUS 3
MOS 64KB
CPU MUMPS INTERP
SYNCHRONOUS 19.2/50K BAUD
OPERATOR'S CONSOLE
40 MB DISK
HARRIS SLASH/6 VM 48KW
PROGRAMMER PANEL
OCR SCANNER
SCIENTIFIC & ARITHMETIC UNIT

The prototype can be broken down into three functional parts which have been named for convenience the "front-end," "middle," and "back-end" subsystems. The Artronix Modulex, or a similar system, will comprise the middle. The actual applications code which supports the administration of test items will, in large part, be contained in the middle system, and most file maintenance will be performed there. The back-end system will closely resemble the Harris/6, a virtual memory computer with a 24-bit word size and a parallel scientific arithmetic processor unit. The Harris/6 will be responsible for the number crunching, notably floating point number crunching, and array processing associated with adaptive testing, assignment algorithms (if implemented in the mobile environment for recruiters) and career information analysis. In the front end will be located several microcomputers which are responsible for handling the graphics display information, decoding, transmission of limited-keyboard information, polling of many devices, and selection of appropriate station or information received by it from the other subsystems.

Other configurations under consideration are shown in Figures 3 and 4. These configurations represent differing emphases upon the I/O or computing requirements of a live testing environment. Obviously, a final selection cannot be made without considerable further study of such areas as resource utilization, component reliability, and cost trends.

Figure 3
Configuration B

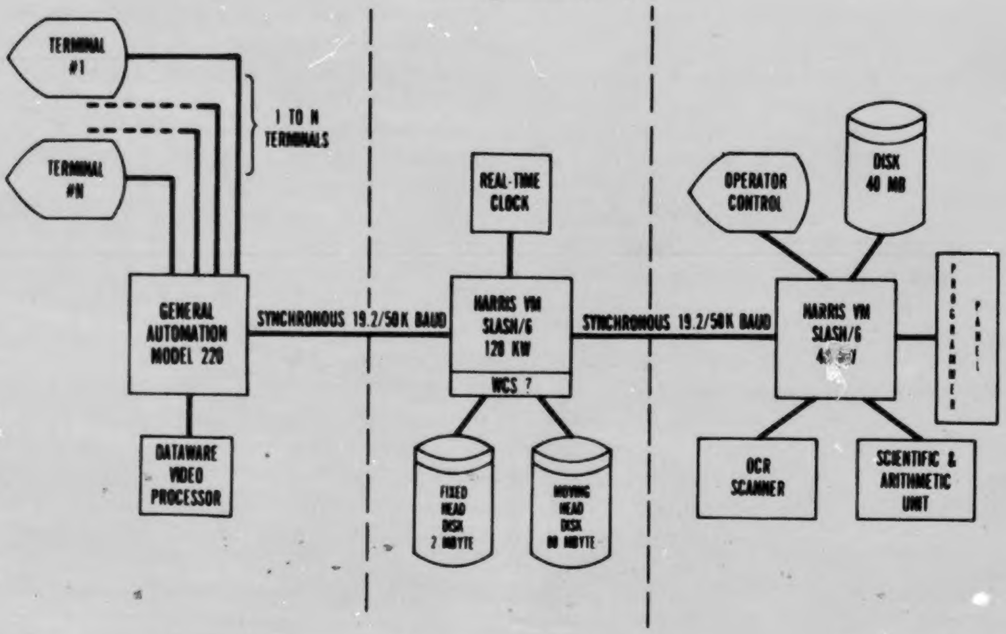

TERMINAL #1
1 TO N TERMINALS
TERMINAL #N
DATAWARE VIDEO PROCESSOR
MICRO-PROCESSOR
RS232C 9600 BAUD
REAL-TIME CLOCK
HARRIS VM SLASH/6 128 KW
WCS ?
SYNCHRONOUS 19.2/50K BAUD
OPERATORS CONSOLE
DISK 40 MB
HARRIS VM SLASH/6 48 KW
PROGRAMMER PANEL
DATAWARE VIDEO PROCESSOR
MICRO-PROCESSOR
FIXED HEAD DISK 2 MBYTE
MOVING HEAD DISK 80 MBYTE
OCR SCANNER
SCIENTIFIC & ARITHMETIC UNIT
TERMINAL #1
1 TO N TERMINALS
TERMINAL #N

Considerations in Testing System Design

Item Presentation

Those who have considered adaptive administration of military tests may consider the Armed Services Vocational Aptitude Battery (ASVAB) as one possible item pool source. With this in mind, the items in the ASVAB were examined to determine the computer resource requirements of item presentation. Resource utilization occurs with respect to item presentation along several dimensions: (1) storage, (2) display formats, (3) system response time, (4) number of examinees to be supported, and (5) examinee response measurements. Economies along these dimensions can directly affect test construction and evaluation strategies, as well as the cost and performance of the supporting computer configuration.

Figure 4
Configuration C


TERMINAL #1
1 TO N TERMINALS
TERMINAL #N
GENERAL AUTOMATION MODEL 220
DATAWARE VIDEO PROCESSOR
SYNCHRONOUS 19.2/50K BAUD
REAL-TIME CLOCK
HARRIS VM SLASH/6 128 KW
WCS ?
FIXED HEAD DISK 2 MBYTE
MOVING HEAD DISK 80 MBYTE
SYNCHRONOUS 19.2/50K BAUD
OPERATOR CONTROL
DISK 40 MB
HARRIS VM SLASH/6
PROGRAMMER PANEL
OCR SCANNER
SCIENTIFIC & ARITHMETIC UNIT

"Front-end" clusters. The best estimates and evidence of the Data/Ware 12/40 image-processing system show that a high-speed microprocessor is capable of expanding compressed video images and displaying them at as fast a rate as 7.5 256 X 256 dots or "pixels" per second. This clearly demonstrates the feasibility of providing high-performance graphical item presentation for a large number of terminal-type devices with a single microprocessor. As the performance information about adaptive test image presentation becomes available from prototype systems, estimates can be made concerning the number of devices that can be supported by a given hardware configuration. The number which can be supported is a function of the available memory for data areas in the "middle" subsystem, the amount of floating point arithmetic and array processing which must be performed, the percentage of graphical items presented, the rate of progression through the test by respondents, and the speed of communications with the middle and back-end subsystems. Additional overhead must be allowed for performing poll-and-select operations upon data coming to and from the terminal-type devices or the test stations.

Storage. Moving head disk drive and controller assemblies on mini-computers cost between $8K and $40K, depending upon total storage capacity and rated transfer speed. At a cost of about .045¢/byte, or .0057¢/bit, low-speed storage on high density drives may seem relatively inexpensive when compared to floppy disk drives, for instance, at .64¢/byte. However, the expense of such drives on mobile systems represents as much as the cost of the CPU and a considerable amount of main memory. The requirements for storage must be carefully estimated and monitored while the system is in operation. Storage of item information consists of item parameter data and the item itself; the storage needs of the latter are of particular interest.

Existing computer terminal technology provides for increasingly lower-cost display of standard 5X7, 7X9, or better dot-matrix alphanumeric symbols. While the low end of the cost per unit with industry standard bare bones features seems to be around $850/unit, this is not likely to decrease, although the number of available features may increase due to increased use of microprocessor-based terminal designs.

This technology would handle roughly 60-75% of the existing ASVAB item pool and with the use of intelligent compression strategies for textual data could occupy as little as 480 bytes times the number of items (200 items = 96,000 bytes), or about $43 of disk storage excluding system overhead. However, increased item pool sizes would be desirable for such reasons as security, increased test resolution, and sophistication. It seems to be not unreasonable to assume the need for a very large item pool for the purpose of uncompromisable, comprehensive ability assessment for military personnel accessioning. The storage would range from 1.2MB to 2.4MB and cost around $1,080 for the large drives. Utilization of available space would be at about 3.5%. This would rule out the use of floppy disk drives and smaller cartridge drives.

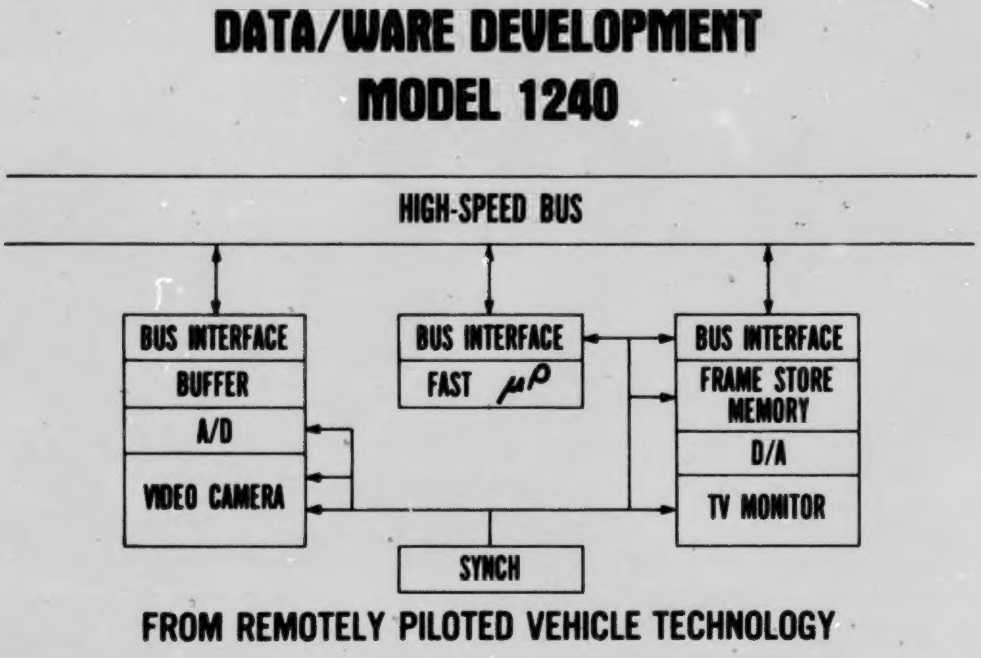
The test administration programs themselves would not require more than 1.5MB storage, including dynamic data areas. One adaptive testing program studied utilized only 16 bytes per item for psychometric information (item difficulty, discrimination, guessing coefficient); and this information would consume only another 40-80KB. However, it may be desirable to store additional item information at this level for on-line use of graded response models (Samejima, 1975).

Display formats. Technology is available for presenting conventional alphanumeric information on a CRT, and the cost per unit could be reduced by minimizing capabilities. However, the presentation of pictorial items is another matter. Plasma, microfilm, and conventional graphics terminals presently cost between $5,000 and $80,000 depending upon capabilities. At $5,000 per station, adaptive testing could never become cost-effective. Therefore, an investigation of military research and development into image processing techniques was made. The strategy used in remotely piloted vehicles seems to hold the most promise. This technique involves (1) the use of high-speed microprocessors and image compression methods to convert analog conventional video signals to a digital format for encoding for either storage or analysis by computer; (2) transmission by telemetry in digital form; and then (3) ground reconstruction back to a video format through decoding of the digital signal and the use of a video gray scale.

The technique will be studied for applicability to adaptive testing by a pending contract with NPRDC. The appeal of this approach is manifold: (1) it can reduce the cost of graphics-capable stations to well under $1000/station; (2) it provides for a built-in image compression scheme which will conserve storage of items in the system; and (3) it gives psychologists better control of the stimulus configuration by providing it in a computer-manipulatable format. Naturally, it will be possible to superimpose alphanumeric and graphical information upon a single image. Figure 5 outlines the way a currently operational video-based graphics system works as developed by Data/Ware Development under contract to NOSC and Wright-Patterson AFB.

Even utilizing dynamic compression techniques, storage of graphics items for adaptive test administration requires considerable storage on disk. It has been estimated that a typical item would require as much as 50K bits, depending upon the nature of the image(s). This would require at least another 1.5MB on disk. Further, because of the availability of this capability, the number of graphic items may increase for adaptive testing applications. Not nearly enough is known about the parameters of such processing, and the pending contract will address these issues in a more quantitative fashion, as well as setting out hardware and software specifications for a prototype system with these capabilities.

Figure 5
System Block Diagram of Data/Ware Development
Video Graphics System


DATA/WARE DEVELOPMENT
MODEL 1240
HIGH-SPEED BUS
BUS INTERFACE
BUFFER
A/D
VIDEO CAMERA
BUS INTERFACE
FAST μP
BUS INTERFACE
FRAME STORE MEMORY
D/A
TV MONITOR
SYNCH
FROM REMOTELY PILOTED VEHICLE TECHNOLOGY

System Response Time

The penalty for waiting for the computer to respond to a command varies, depending upon such factors as the status of the person waiting and the degree of urgency for the particular task. In the adaptive testing situation, waiting for the system has serious consequences. The most obvious effect is that it increases test administration time, the minimization of which is supposed to be one of the advantages of adaptive testing. Less obvious are such factors as the influence upon examinee attention, the significance of environmental distractors, and anxiety and motivational factors. System response time is also an important dimension in a small computer system; it is even more important because of the relative ease with which applications can bog down the system. In general, systems programming and applications programming must be carefully inspected; and all jobs must be of a carefully predetermined and carefully engineered variety. It is desirable for the job mix, as much as possible, to remain relatively homogenous with respect to resource utilization and program behavior.

Many factors can cause decreases in response time. Probably the most important to keep in mind is the small number of resources. For instance, there is only one disk for general system use; and contention for it will be high, especially if the amount of main memory is kept small for economy purposes.

Number of Examinees to be Supported

In determining system cost-effectiveness, the number of simultaneous test administrations which can be performed is a very important factor. In general, with minicomputers as the number of simultaneous users increases, the cost of the overall configuration rises in direct proportion. Analysis has shown that conventional channel multiplexing techniques for data communications (usually RS232C ports) would be an inefficient way of communicating with a large number of test stations. A design objective has been to develop a capability for as many as 64 simultaneous administrations. Whether or not this is feasible with the contemplated design is difficult to answer at this time. Table 1 shows some of the parameters influencing the number of stations which can be serviced in a reasonable response time. Sixteen administrations is a lower bound for system cost-effectiveness, i.e., simultaneous administration of fewer examinees would reduce system cost-effectiveness to an unacceptable level. While the AFEES environment may not presently require so many, a high-school administration would have to occur on a very large scale in a relatively short period of time in order to compete with the economy of scale offered by conventional paper-and-pencil tests.

System costs can be broken into fixed and variable cost factors. In order to administer only 16 tests, a certain minimum system is required. As the number of examinees increases, memory and processor use are utilized in direct proportion to the number of administrations and the amount of disk concentration goes up by a more complicated formula. Cost-effectiveness assessment for this kind of configuration is difficult to perform; and it is necessary to know data external to the computer system performance, such

Table 1
Factors Influencing Number of Serviceable Stations

Speed of Dataware Video Processor
I/O Capacity of Microprocessor
Main Memory Size of Microprocessor
Speed of I/O Interface from Front End
I/O Buffer Size of Front End
I/O Handling Speed of I/O CPU
Space and Electrical
Back-end Computing Speed
Size of Code and Data Areas for all Programs

as average test length under adaptive testing in order to arrive at better estimates. The value of field tests in this regard cannot be exaggerated. The desirability of large-scale administrations also depends upon policy factors and the working relationship with public institutions for which testing may be offered as a diagnostic and/or vocational service. Although direct comparison with the present ASVAB program is tempting, the quantum step forward represented not only by the computer system itself, but also by the adaptive testing strategy, suggests that comparison would be inappropriate.

Examinee Response Measurements

The adaptive test program studied utilizes a dichotomous (correct/incorrect) response structure. There are other response measurements which can be made. For instance, time-to-respond, or latency, could be collected. For graphic items, relative image size and resolution may be relevant; an individual "calibration" might be required, based on the conclusions of basic research into the relationship between stimulus characteristics and examinee performance on an item.

Indices of Testing Cost-Effectiveness

Issues of cost-effectiveness will be raised more than once before adaptive testing is widespread in its application. Contributing to the confusion about the cost of adaptive testing will be the fact that hardware (i.e., primarily computer hardware) is involved in adaptive testing, whereas conventional paper-and-pencil testing is essentially labor-intensive. Conventional testing is costly in terms of examinee and proctor hours and scheduling of space and people; however, these items are not directly chargeable and represent consumables rather than capital investments. In a technologically advanced culture such as ours, it is relatively more straightforward to develop algorithms for justifying hardware investments where labor savings can be shown. The ease with which automated testing techniques can be "sold" will depend upon the consequences of the current inconveniences, inaccuracies, or difficulties associated with paper-and-pencil testing. In military personnel acquisitions, the problem is also one of market penetration: to test as many people as possible in as short a period of time to minimize disruption to the host institution's environment. Some of the parameters which influence adaptive testing cost-effectiveness are:

1. Mean response time of examinees to (video-presented) items;
2. Computer system configuration size;
3. Mobile support subsystems cost (if applicable);
4. Maintenance costs for hardware and software;
5. Training associated with staffing of computerized adaptive testing stations or vans;
6. Number of items required to perform necessary ability measurements;
7. Capability of computer system to perform other recruitment-related tasks (e.g., computerized career information systems, data management, individual job assignment), with associated cost-reducing effects;
8. Cost and speed of data communications (very little is involved in the configurations proposed in Project CONTRACT);
9. Number of hardware items procured within an 18-month time frame;
10. Estimated system life for hardware and software;
11. Installation costs for testing stations;
12. Research and development: computer science, psychology, operations research;
13. Hardware and software quality and reliability;
14. Number of simultaneous test administrations; and
15. Frequency and cost of item calibrations.

A sample systems cost is presented in Table 2. Relevant indices are cost per item administration, cost per test, number of simultaneous administrations permitted, and daily operating expense. Obviously, the number of items required to make ability estimates and the complexity of testing algorithms employed are primary factors in the determination of cost effectiveness.

Problem Areas

Systems analysis of the configuration alternatives for adaptive testing and of the software techniques presently in use for adaptive testing have suggested some problem areas that are worthy of attention. This attention must be given jointly by computer scientists and psychometricians if a cost-effective, fully functioning system can be delivered in a reasonable time-frame. At best, topic areas can be delineated in the present discussion; further research is obviously needed.

Frequency of estimate update. Estimated mean and variance of a normal Bayes prior distribution, given observation of dichotomous response to an item with known characteristics, can be updated on an item-by-item basis or after a set of items have been answered. CPU (though not I/O) demands could be minimized through the use of item sets where possible. Item set size could be a function of the amount of precision required in the ability assessment, logical or substantive interrelation, item structural similarity, or simply a convenient "page" size for use on a video screen.

Table 2
Sample Systems Cost (Estimated)

Item	Cost
Computer System @ 50 unit quantities Includes: 5 independent memory "banks" Six minicomputers 110 MB moving head disk storage Matrix printer (500 cps) OCR Front end microprocessors	210K
Video Devices (64)	32K
Van Excluding Power and Air Conditioning	30K
Power and Air Conditioning	10K
Total	282K
*Maintenance (.5% of purchase per month, serviced by government personnel)	700/mo.
*Transportation (10K miles per year, vehicle maintenance, etc.)	300/mo.
*Communications	45/mo.
*Training and Miscellaneous Software Maintenance	100/mo.
5-year Total Recurring Costs	68.7K/5yr.
Installation	2K
Assumptions: 64 stations, 70% overall utilization; School calendar (180 days) in use; 35 items to criterion for adaptive testing; Item presented average of every 20 seconds;	
Over 5 years, could administer: 1,240,000 test @ 28¢ per test**	

* Probably would be offset by manpower savings through improved procedures in recruiting.

**"Test" assumed to be one sequence of adaptively administered items whose mean length is 35 items.

Pool size. An important topic is the implication of item pool size upon test reliability, precision, security, and administrative efficacy. Pool size affects the storage requirements and, if it reflects a more sophisticated item-sequencing algorithm, increased CPU usage. Items can have different "versions" or possess redundancy along substantive dimensions. A data base structure which reflects such relationships must be designed to allow for intelligent access to item raw data by applications programs.

Precision. One problem with many small computer systems is that they are not designed for high-precision applications. This is frequently a function of the common 16-bit word size; even double precision floating point arithmetic can result in insufficient computational accuracy, perpetuation of roundoff errors, truncation, and loss of significant digits. This problem can be exacerbated if the computing for sequencing moves at all in the direction of a convergence failure. A parallel-processing floating point array processor may be indicated if conventional mini-architectures are not adequate to keep up with the demands of this applications area. Also, given that parity errors do occur in disk and main memory systems and because losses of single bits can have disastrous results in the computation of statistical algorithms, this must be provided for in advance. Hardware must be specified which has main memory byte parity and correcting and checking.

Storage of sequence information. As each examinee traverses through an adaptively administered test, a probable unique sequence pattern is generated. This pattern is a form of response data which may or may not be utilized by the adaptive algorithm for administering the next item(s). As this sequence grows lengthy, the overhead associated with maintaining that information and incorporating it into analysis becomes significant. The tradeoff between the precision gained by such sequencing data, if any, and the associated computing overhead should be estimated; such sequencing data may have some value in evaluation of the algorithm and the testing process as a whole.

Possible use of differential starting values. It has been suggested that externally determined criteria for the starting point in an adaptive test administration may be utilized. This significant possibility raises the idea of an integration of the career information aspect and the testing aspect of the mobile computing system for recruitment. Counseling-type information may suggest starting points. Other test scores may suggest starting points, or recruiters might make use of the adaptive test capability to perform screening (in which administrations would not be carried to completion). The interrelationship of programs and data bases for these various application areas must be carefully studied to determine if dynamic task synchronization should be made another requirement of the target system's operating system software.

Use of FORTRAN. The acceptance of FORTRAN for the implementation of adaptive testing algorithms has both advantages and disadvantages. A prominent advantage has been that psychometricians have been able to write some or all of the code. For the implementation of statistical algorithms, excluding data manipulation, FORTRAN is adequate and efficient on small computer systems. However, its use for any I/O--especially disk I/O--is contraindicated. This can be said for most of the FORTRAN implementations on small computer systems. FORTRAN is especially poor for formatted Input/Output because it uses excessive CPU resources and makes poor use of memory while doing it. Access to complex file structures, too, may be cumbersome. The use of FORTRAN must be restricted to those domains in which it generates efficient and controllable code for small computer systems. Certainly, only part of the adaptive test administration process can utilize FORTRAN, for its reckless use can result in needless proliferation of CPUs and memory.

Requirements of Adaptive Testing Systems

Program Attributes

An important means of determining appropriate hardware and software requirements for adaptive testing is to study the characteristics of existing programs. Examination of these FORTRAN programs reveals useful information which is presented in Table 3. Separation of code and data areas by some minicomputer compilers (an important capability) permits direct inspection of the code and data requirements for each routine, including shared COMMON data areas. Program code, or instruction space, is the only program memory resource which can be shared between multiple executing users (or, in this case, examinees). Therefore, the ratio of code to data memory resources needed by these programs is an important characteristic. In the item administration program summarized in Table 3, a maximum of 24% of the memory resources needed can be shared between users. In a minicomputer environment with a target support capability of 64 simultaneous users, this is unacceptable; a total of 749,312 bytes would be needed for data areas in the worst instances.

Table 3
Program Attributes of a Bayesian Ability
Estimation Program (Simulation of Actual Sessions)

User Code Segments: One Main, Two Subroutines
1. 2364 bytes
2. 165
3. 252
4. (FTN Intrinsics) 902

TOTAL: 3683

User Data Segments
1. 10755
2. 72
3. 80

TOTAL: 10907

Ratio of Code/Data Areas for Program as a Whole:

3683/10907 = .338
Data Areas Represent 74.8%

File Space (if all files open): 801 bytes
Minimum Memory Required by B1700: 3643
Dynamic Area Required for all Data Pages to be in Memory: 14216
Percentage of Sharable Memory Space = 23.9%
Execution Time on B1700 for 32 Simulated Subjects with Maximum 64KB Main Memory, and 16 Items = 5 min. 12 sec.

On the NPRDC minicomputer, a 64KB Burroughs B1700, the operating system's virtual memory capabilities make it possible to execute programs for which the data and/or code space requirements are greater than what is actually available on the computer system. This is accomplished by breaking up the code and data areas into variable size "pages," which can be written to disk and then brought into memory when needed and/or when space becomes available. The Burroughs B1700 FORTRAN compiler identifies at BIND time its minimum dynamic memory requirements. For the item administration program, only 3643 bytes of the total required 15391 data, code, and buffer areas need be present in memory for the program to execute. However, the penalty to pay for this capability is the overhead associated with monitoring page references, reading and writing pages to disk, and producing compiler output that is pageable. The generally slow speeds associated with these techniques are not adequate to the task of high-performance test administration/adaptive testing while simultaneously performing counseling and other application programs. The only solution is to improve the way in which the programs are written and to configure hardware that is exceptionally well-suited to the job at hand. Table 4 shows some common programming flaws in the software studied. As can be seen, a number of poor programming techniques which can severely reduce testing throughput time are evident.

Table 4

Some Poor Programming Techniques Found in Item Administration Programs

Every iteration of some routines recompute items that could be set in DATA or computed just once at execution time, for example :

```
    PI = 3.14, etc.
    C1 = 1.0/PI
    C2 = 2.0/SQRT(PI)
    SIGN = -1.0
```

No use of tabbing on formatted output

Excessive use of FORTRAN formatted I/O (very CPU-consumptive)

Excessive use of GO TO's; unstructured logic is common; makes it difficult to optimize code

Inefficient array referencing and dimensioning

Successive IF statements not placed according to frequency of branch (where applicable)

Space-wasteful ordering of variables in BLANK COMMON

Other Requirements

Use of firmware. The advent of user-microprogrammable minicomputers (e.g., those by General Automation, Hewlett-Packard, Varian, and Computer Automation) has made it possible for particular repetitively executed primitives, which are very low-level software constructs, to be coded in firmware. Some of these firmware capabilities are evidenced in the Hewlett-Packard 21MX mini on which a number of FORTRAN-generated primitives are available in microcode. Possible areas for microcoding for the adaptive testing algorithm are: absolute value function, min and max, float, fix, Pearson's *r*, modulus, passage of subroutine arguments, and multiple dimension array references.

Hard copy output. It is not clear at this point what the requirement for hard copy output at the end of an adaptively administered examination might be. As the need for high-speed hard copy could represent a considerable investment, some thought should be given to whether this is a requirement, and if so, how fast hard copy must be produced and in what form.

Test calibration and propagation. This treatment of computer requirements for adaptive testing has not addressed the issue of test calibration, which is currently performed primarily on large scale computer systems. A nationally utilized test could need field updates or new calibration data. The frequency of such update should be estimated. It should also be determined what data should be collected by field systems to aid in refinement of instrument precision. The use of the computer for test administration should aid in continual improvement in the test merely through simplifying data analysis procedures and facilitating field implementations of new versions of the test pool or algorithms.

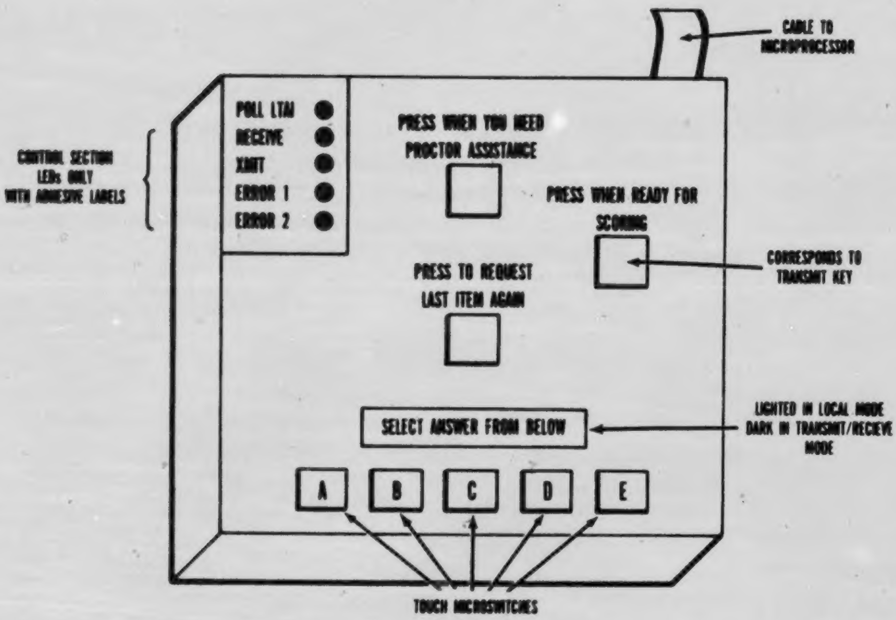
Recovery, reliability,and cross-checking. The possibility of catastrophic failure always exists in non-redundant computer systems. Total hardware redundancy for the system under consideration was not judged to be a cost-effective additional capability. Therefore, recovery from software (either caused by systems software or applications software failures) is a contingency which must be planned for. Complete automated test data generators should be used (such software is already available on larger computer systems and could be adapted for use on small systems) to exercise the software fully before it is pressed into full use. Whenever possible, software should have built-in cross-checks to assess its proper functioning. A valuable heuristic approach is to assume that physical system security has been compromised. File protection (i.e., protection of data and programs) is as important as physical system protection. Simple errors such as channel cross-talk could result in mix-up of item presentations with responses; this is a phenomenon sometimes observed in many timesharing systems when pointers are lost or files are partially destroyed.

Recovery from hardware failures is also significant. The use of periodic "dumps" is worthwhile so that examinees do not have to totally restart examinations. However, there is overhead associated with all of these precautions; and their cost must be weighted against the value of the protection.

Keyboard design. It is doubtful that a full typewriter keyboard is required for adaptive test administration. The technology developed by the Air Force Human Resources Laboratory (Kirby & Gardner, 1976) shows attention to cost and functional requirements along these lines. Psychologists and engineers must plan a minimum keyboard design which is reliable, easy to use, self-evident in function, and totally detachable for plug-in replacement. The way has been shown by the financial institutions, as special teller keyboards have been designed to meet specific transactional requirements. Some inherent logic should be present, such as illumination of error conditions and of line activity to prevent key-in attempts during transmission while buffers are being filled.

Figure 6 shows a possible minimal keyboard for an adaptive test station. It is generally accepted that five-alternative multiple-choice items work better than those with fewer alternatives. For this reason, the suggested keyboard interface layout is shown with a five-response touch microswitch capability.

Figure 6
Possible Minimal Keyboard for
Adaptive Test Station


CABLE TO MICROPROCESSOR
CONTROL SECTION LEDs ONLY WITH ADHESIVE LABELS
POLL LTM
RECEIVE
XMIT
ERROR 1
ERROR 2
PRESS WHEN YOU NEED PROCTOR ASSISTANCE
PRESS WHEN READY FOR SCORING
CORRESPONDS TO TRANSMIT KEY
PRESS TO REQUEST LAST ITEM AGAIN
SELECT ANSWER FROM BELOW
LIGHTED IN LOCAL MODE DARK IN TRANSMIT/RECIEVE MODE
A
B
C
D
E
TOUCH MICROSWITCHES

Computer Technology and the Stimulus-Response Configuration

Some experts have speculated that the future holds a considerable number of alternatives for departure from typical stimulus and response configurations for test administration. The role which the computer and related peripherals can play in contributing to the increased comprehensiveness of psychomotor, ability, perceptual, and behavioral assessment depends to a large extent upon the extent of dramatic cost-reductions in many electronic components and the strides made in software engineering. The rate of these transformations has not been as great as had been thought. The cost of sophisticated graphics, or image manipulation devices, ranges between $15K and $200K per station. This is certainly not appropriate for individual adaptive test stations. So far as natural language processing is concerned, speech understanding systems lag considerably behind speech synthesis. The computer is very proficient at cost-effective presentation of alphanumeric information, so that presentation of spoken instructions or test items would have to be justified on an affective, motivational basis or increased psychometric sensitivity, which is difficult to make cost effective.

The U.S. Civil Service Commission has stated that a nationwide computer-assisted testing system for federal jobs is being developed for the 1980 timeframe (Urry, 1977). The design of a nationwide network should be undertaken with a serious view to the level of processor which is required to support the chosen adaptive testing strategy and the cost of a network, if one is needed. The systems design NPRDC has undertaken to identify a system capable of the implementation of sophisticated adaptive testing algorithms does not require a nationwide network in a real-time sense of a network. The costs of such a network are extremely high, simply for communications costs alone. A centralized system design is less likely to be cost-competitive. Furthermore, the hope for instantaneous and continual application of extremely complex theoretical/statistical procedures must be tempered by the high probability of needing a costly computer configuration to support them.

Conclusions

It has been the intention of this discussion to raise more questions than it has answered. This is the case because of the need for considerable research involving psychometricians and computer scientists in developing a small computer system equipped with the proper combination of hardware and software architectures, which is priced low enough to make adaptive testing cost effective and with enough power to get the job done simultaneously for a sizeable number of examinees. The use of a prototype computer system is essential even if not all of the possible alternatives can be quantitatively evaluated prior to its procurement. With such a prototype, experiments can be designed, simulations can be performed, and hardware and software configurations can be benchmarked. At this point, the questions are numerous enough so that it is difficult to order them according to their impact on system performance. Despite these reservations, the technology of the minicomputer has come a long way in recent years; and much more is possible with the same hardware than was previously. Determining where off-the-shelf hardware and software is not adequate so that customized engineering development can be initiated is one of the first goals, because of the lead time required for development prior to deployment.

The policy and applications contexts into which the adaptive testing system will be placed are also important systems design parameters. It is generally agreed that a multi-purpose mobile system has broader appeal and greater cost-effectiveness justification than one dedicated solely to testing. Where testing is to be part of a larger computing environment, its resource requirements and program behaviors must be assessed in the complex of the overall system control, especially control of memory allocation. If the need for a comprehensive information-dispensing and information-organizing capability is not immediately apparent, recruiters, with their necessarily urgent needs for meaningful and marketable data, will certainly make known the need for a fully integrated system. The parts to be integrated will include: career policy guidelines, information, some goaling data, a mini job-assignment capability, adaptive testing, logging of psychometric data, and recruiter management information. If such a system were to be delivered next year, there can be little doubt that this list would grow simply because of the added capability that a computerized system represents.

References

Kirby, P., & Gardner, E. Microcomputer controlled, interactive testing terminal development (AFHRL-TR-76-66). Lowry Air Force Base, CO: Air Force Human Resources Laboratory, Technical Training Division, 1976.

Samejima, F. Graded response model of the latent trait theory and tailored testing. In C. L. Clark (Ed.), Proceedings of the First Conference on Computerized Adaptive Testing, Washington, DC, 1975, pp. 5-15.

Urry, V. Tailored testing: A successful application of latent trait theory. Journal of Educational Measurement, 1977, 14, 181-196.

Acknowledgements

The opinions expressed are solely those of the author and do not necessarily reflect official Department of the Navy policy.

IMPLEMENTATION OF A MODEL ADAPTIVE TESTING SYSTEM AT AN ARMED FORCES ENTRANCE AND EXAMINATION STATION

MALCOLM JAMES REE
PERSONNEL RESEARCH DIVISION
AIR FORCE HUMAN RESOURCES LABORATORY
BROOKS AIR FORCE BASE, TEXAS

In a world of increasing technical complexity and diminishing resources, it is the task of the military recruiting agencies to obtain the most highly qualified candidates for technical training. Traditionally, paper-and-pencil multiple-aptitude test batteries have been administered to applicants of a wide range of abilities. These tests have been peaked to be most discriminating over a relatively narrow range, because limited time precluded administration of enough items to gain maximal test information over the broad range of an ability. Selection and classification decisions must be made, however, which require discriminations at the 80th percentile. At this level, only limited information is available from a typical peaked test.

Adaptive testing, particularly computer-driven adaptive testing, promises (1) to enable the gathering of test information (Lord & Novick, 1968, Chap. 20) at all levels of ability with equal precision and (2) to increase the predictive validity of military accession testing. Furthermore, adaptive testing promises to reduce the time required to obtain ability estimates for applicants; possibly, by making accession a one-day process, it may reduce overall costs.

The model adaptive testing system was implemented in an Armed Forces Entrance and Examination Station (AFEES) in order to study its feasibility for use in a military selection setting. At the AFEES, the testing system must (1) be operated by individuals without any special training in computer hardware or software, (2) perform when needed, (3) be operational for the entire workday, (4) accommodate applicants for military service from very low ability to very high ability, (5) not intimidate or frighten the applicants or the test administrators, and (6) provide valid and reliable measurement.

Prior to the implementation of an adaptive testing system, a number of decisions--both technical and administrative--must be made. The technical questions include: (1) who are the subjects; (2) what ability areas are to be tested; (3) what items and item statistics are available; (4) which scoring method will be used; (5) which item selection technique will be used; (6) what media for question presentation will be used; and (7) how will pictorial items be presented. There are also many administrative questions. How can the operation be simplified so that low ability, careless, or inattentive examinees do not cause an abnormal termination of testing? What impact will the demonstration have on day-to-day AFEES operations?

Implementation

The Setting

The San Antonio, Texas, AFEES was chosen as the test site because it was close to the development center at the Air Force Human Resources Laboratory (AFHRL). This proximity afforded considerable opportunity for monitoring the progress of the adaptive testing system.

The subjects for this demonstration were applicants for military enlistment, and their abilities covered a very broad range of aptitudes. They were tested in three aptitude areas which comprise the Armed Forces Qualification Test (AFQT): Word Knowledge (WK), Arithmetic Reasoning (AR), and Space Perception (SP). The AFQT is used for initial qualification for military service. Other aptitude areas are usually measured only if an acceptable score on the AFQT is achieved; these subjects were tested while awaiting the results of the AFQT.

Test Items

The items used for this model adaptive testing system were culled from existing historic item files at the AFHRL. Only item difficulty (p) and item discrimination (phi) indices were available. Items were selected to represent a generally rectangular distribution of difficulties from about .2 to about .8, with the highest available discrimination index at each difficulty level. These items were then assembled into booklets for administration to Air Force basic recruits in order to estimate the latent trait parameters a, b, and c (Lord & Novick, 1968) for later phases of this demonstration. Initially, the classical item indices were transformed via approximations and used to calculate the latent trait parameters useful for the project. Although these estimates varied somewhat from more exact estimates obtained from the new response data, they did permit a reasonable starting point from which to demonstrate the feasibility of adaptive testing for military service applicants. As soon as a satisfactory sample has been collected and the parameters estimated, the approximated parameters will be replaced by the new, more exact parameters.

Computer System

A medium size computer (IBM 360/65) was available for the demonstration in a time-shared mode. The APL programming language (Gilman & Rose, 1970) was selected because it is interactive and has extremely powerful operators. In addition, experience has shown that APL leads to fast program development. It is also fast in execution and is particularly suited for handling vectors and matrices.

A combination of Bayesian item scoring and ability estimation (Owen, 1969) and selection of items by maximum information (Lord & Novick, 1968, Eq. 20.4.1) was selected for ease of programming and low computer core utilization. Two criteria for the termination of item administration are (1) the reduction of the posterior variance of the ability estimate to a low value (<.0625) and/or (2) the subjects having taken 20 items. This procedure is also advantageous because it does not require a structured item pool as would a stratified adaptive test; thus implementation of the testing system is made easier.

A modified Tektronix model 4006-1 Cathode Ray Tube (CRT) terminal was used for the demonstration. Both a viewing hood to reduce glare and a keyboard cover to prohibit pushing inappropriate keys were fabricated. This terminal used the Tektronix Graphics Package, APLgraph 2, and was run at 1200 baud in half-duplex mode.

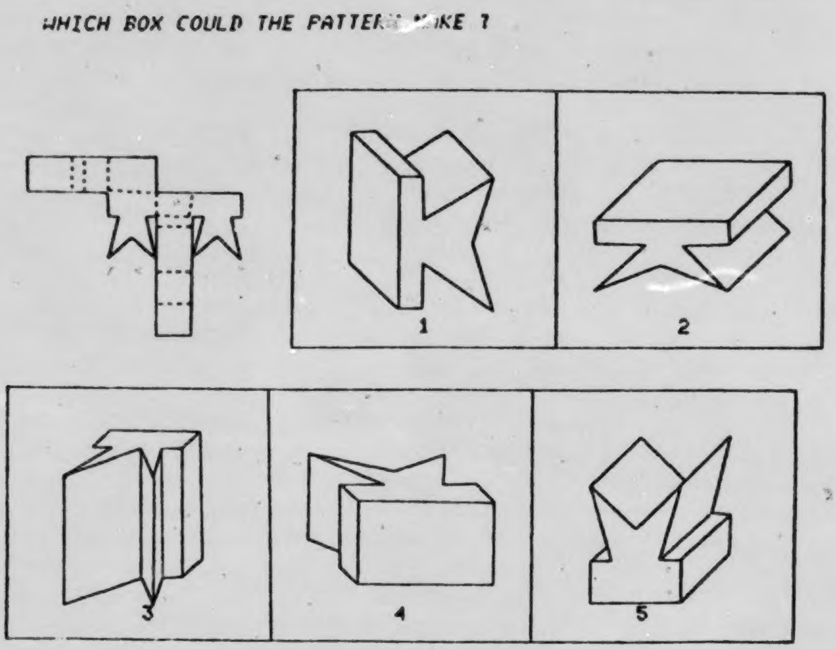
In order to insure proper operation of the system, operating instructions and operating safeguards were built in. The examinee is taught how to use and respond to the terminal before any questions are presented. All solicitations for input are for the characters "1," "2," "3," "4," "5 " and are checked to determine the presence of alphabetic (e.g., ABCD) or special characters (e.g., $!&). If an out-of-range response, an alphabetic character, or a special character is given, the instructions for responding are repeated. Then the screen is cleared and finally the question is repeated. Proper character input is then converted to its equivalent numerical form and processed.

The characters for the questions of WK and AR are kept on an external randomly accessible file and read as needed. Screen control characters are stored with the literal characters, which makes for simplicity of operation. The last array of the file, roughly equivalent to the last record of a FORTRAN file, contains a 4 by N (N = the number of items in the file) matrix of the item parameters and answer keys. This matrix is read in and manipulated prior to the presentation of all questions.

Producing the pictorial displays for the SP items presented a unique problem in storage and drawing. One proposed storage method was to use back screen projection from a random access slide projector. This was discounted because it allowed only about 100 items to be stored, added a mechanical component to maintain, and required photographic slides of each item. Similarly, the idea of writing a specific mathematical function to generate each individual figure was dismissed because it required extensive programming for each new item. Finally, a method of display was developed using the sophisticated graphics capability of APL which only requires placing a drawing of the SP item on a "digitizing tablet" and touching various points on the drawing. These are then transformed into a vector for each item, and the graphics package draws the numerous lines represented by the vector at a very rapid rate (see Figure 1). At 1200 baud, the figures almost flash onto the screen as the vector is read in from its place on the file. This technique could be extended to any item requiring drawings, such as mechanical principles or block counting.

All technology and software developed to draw the Space Perception items are general enough to be used in other testing or educational applications. The perspective and three-dimensional effect are very good, and motion for rotating or shifting the figures can be added. Rotating figures, demonstration

Figure 1
Direct Copy of a Typical Space Perception Item From Screen of CRT


WHICH BOX COULD THE PATTERN MAKE ?

of mechanical principles or moving lever arms may lead to new item types not amenable to static paper-and-pencil tests. Computer driven graphics may enable measurement of new and important ability areas.

An operator's manual, algorithmicly written, was produced for the AFEES personnel. It contains complete instructions for initial daily starting and stopping of the testing system; it also gives instructions for starting the program if the terminal is already running. The manual offers names and telephone numbers of people to contact in the event of trouble. The programs have been "locked" to the AFEES staff, and they have been advised not to try to edit the programs. Back-up copies of both the programs and the files are stored on-line and require only a command from the proper user to reinstate damaged programs or to update programs as they are refined.

Data grade telephone lines, a special telephone number for the AFEES use only, and a special sign-on code were provided to reduce competition for telephone ports in the time-shared environment. The "Special Testing Room" at the AFEES used to house the terminal is a 10' x 12' windowless room containing several student chairs with arms, one side chair, and a 3' x 2' table for the terminal. The terminal and the telephone connector need little space and can be operated in any room with 117 volts AC current and a telephone.

Demonstration and Future Implementations

The feasibility of adaptive testing will be investigated in the San Antonio, Texas, AFEES demonstration by the assessment of two important factors. First, did the system run with little trouble and attention? This will be assessed from interviews with the AFEES staff and from daily logs of the system's operation. Secondly, was adaptive testing as valid as paper-and-pencil testing? The validity of the adaptive testing system will be assessed by comparing the subjects' adaptive scores and the subjects' AFQT subtest scores. Analysis of these data will help in making future decisions about adaptive testing.

Before any large scale implementation can be undertaken, there will be questions to answer subsequent to the demonstration. Some of these questions are psychometric, some logistic, and some economic. As yet, no testing configuration, either local or nationwide, has been developed, nor have system costs for implementing, operating, and supporting adaptive testing been established. Basic conceptual questions dealing with such diverse topics as testing models, back-up systems, operating policies, and central versus dispersed processing remain unanswered.

It is conceivable that certain other decisions will facilitate broad scale implementation of adaptive testing. For example, the AFEES in Baltimore, Maryland, already has computer-automated management and paper handling on an in-house mini-computer. The addition of adaptive testing might require little additional hardware; and, in quantity, this additional hardware might be inexpensive enought to merit its use. Furthermore, adaptive testing could add to test security because neither test booklets nor answer key are distributed, and no one can have knowledge beforehand regarding which questions will be administered to a subject.

In the future, the actual costs and benefits of adaptive testing will be known. This will permit realistic decision-making for its use. This knowledge will allow adaptive testing to move from a research topic of the 1970s to an operational tool of the 1980s and beyond.

References

Gilman, L., & Rose, A. APL 360 an interactive approach. New York: John Wiley & Sons, Inc., 1970.

Lord, F., & Novick, M. Statistical theories of mental test scores. Reading, MA: Addison-Wesley Publishing Co., 1968.

Owen, R. A Bayesian approach to tailored testing (Research Bulletin RB 69-92). Princeton, NJ: Educational Testing Service, 1969.

Acknowledgements

The views expressed herein are those of the author and do not necessarily reflect the views of the United States Air Force or the Department of Defense.

COMPUTERIZED ADAPTIVE TESTING WITH A MILITARY POPULATION

STEVEN GORMAN
HEADQUARTERS, U.S. MARINE CORPS

Adaptive testing is a computer-assisted interactive process which facilitates the rapid, accurate measurement of the ability of the testee. The process begins with an examinee, seated at a terminal, being presented a question and responding to that question. After each response, an estimate of the examinee's ability is updated. The computer program then selects a question which is either more difficult or easier. One way to select the next question is to choose, out of a pool of available questions, one that will minimize the standard error of estimate of the examinee's ability when it is administered. This tailoring process, with a minimum number of test items, maximizes the information obtained about an individual's ability level.

With the mathematically elegant Bayesian algorithm developed by Owen (1975) in conjunction with a cathode ray tube and computer interface, adaptive testing is now not only possible, but also necessary for large-scale testing situations, such as military accession testing. This mandate is founded upon the potential benefits of adaptive testing:

1. Greater test precision at all ability levels, especially at the tails of the distribution;
2. Improved test security;
3. Decreased misclassification;
4. Reduction of examinee anxiety or boredom;
5. Reduction in test length;
6. Enhanced discrimination of testee abilities;
7. Enhanced applicant motivation as a result of immediate feedback on test responses;
8. Standardized test administration;
9. Improved data quality through elimination of human requirements for calculation and data recording; and
10. Interface with classification, assignment, and job information systems.

Adaptive testing is based upon latent trait theory--the theory which analyzes examinee performance at the item level. Accordingly, it is extremely important that items be carefully screened for use in adaptive testing and that the item parameters be satisfactorily estimated. The assumptions and procedures for "norming" items in adaptive testing require greater information than was the case with norming tests under classical test theory. These requirements and procedures have been stated previously (Urry, 1976, Jensema, 1976a). A discussion of these requirements and the effects of their violations is in order; this paper will subsequently present a discussion of the item parameterization, the adaptive testing research design, and the metric base for each procedure.

A Brief Overview of Theory

The only appropriate model for representing multiple-choice items is the three-parameter latent trait model (Birnbaum, 1968; Urry, 1971). This model can be represented in either a normal ogive or the more popular logistic function:

$$P'(\theta) = c_i + (1-c_i)P, \qquad [1]$$

where

$$P = \frac{1}{\sqrt{2\pi}} \int_{a_i(\theta-b_i)}^{\infty} e^{\frac{-x^2}{2}} dx \cong \frac{1}{1 + exp[-Da_i(\theta- b_i)]} \qquad [2]$$

a_i = the item discriminating power;

b_i = the inflection point of the item characteristic curve, or the item difficulty;

c_i = the lower asymptote of the regression of item response on the latent trait, also referred to as the guessing parameter;

D = the constant 1.7; and

θ = the latent trait continuum of ability, which ranges from $-\infty$ to $+\infty$, but usually is restricted to the range -3 to +3.

Figure 1 depicts an item characteristic curve where a_i=2.0, b_i=0.0, and c_i=.18. The curve is based on a plot of $P'(\theta)$, the probability of successfully answering a test question, given ability level θ, when guessing is effective. The value c_i=.18 may occur in a multiple-choice test item that has five response alternatives. Note that the c_i value is less than .20. This may be attributable to the attractiveness of the wrong alternatives at greater than chance level. The value b_i=0.0 occurs at the probability location P=.5(1+.18)=.59. The value a_i is related to the slope at the point b_i, which is the inflection point of the curve.

Figure 2 depicts $Q(\theta)$, the curve displaying the probability of having ability θ, given that an examinee responded incorrectly to an item. As can be seen in Figure 2, the slope of $Q(\theta)$ is steeper than $P'(\theta)$, shown in Figure 1.

Figure 1
Item Characteristic Curve (P')

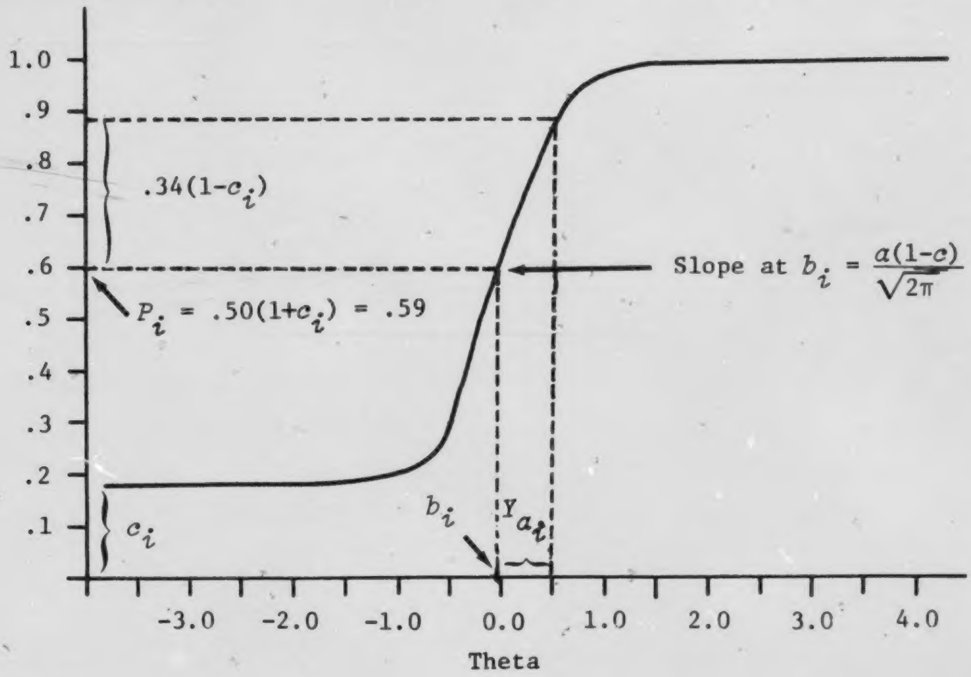
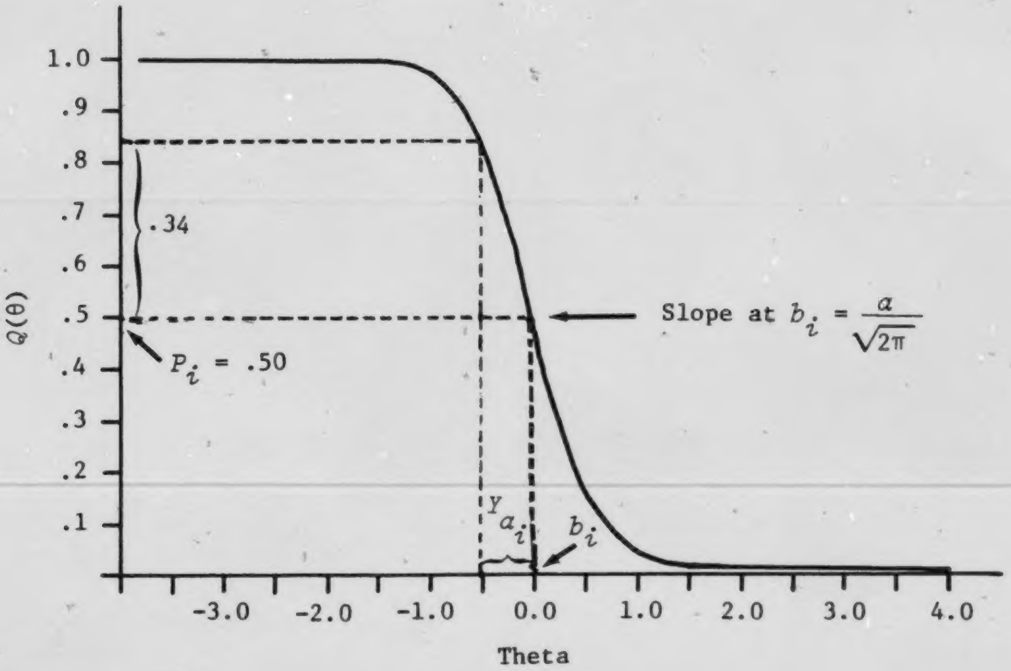


Figure 2
Curve for Wrong Response (Q)



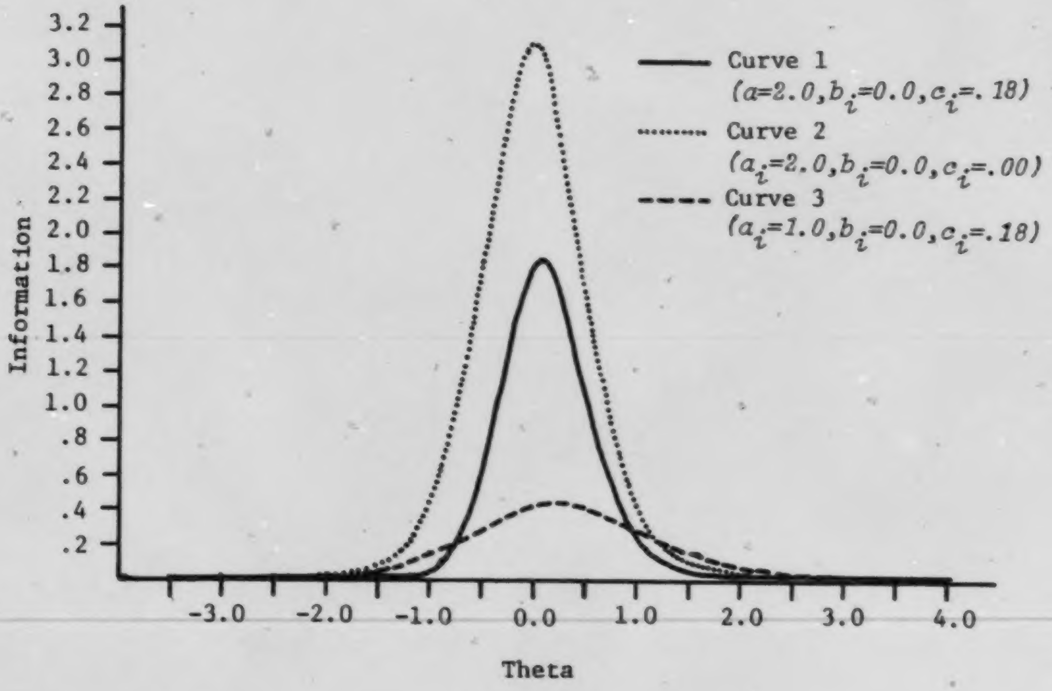
As will be shown more clearly later, an incorrect response to an item decreases the standard error of estimate more rapidly than a correct response.

Birnbaum (1968) developed the concept of information in a test item. The simple formula is

$$I(\theta) = \sum \left(\frac{[\text{slope at } \theta]^2}{\text{variance}} \right) \qquad [3]$$

Figure 3 displays the information available in three test items. Curve 1 is based on the item with a_i=2.00, b_i=0.0, and c_i=.18 which is displayed in Figure 1. Curve 2 has the same a_i and b_i parameters, but c_i=.00. Curve 3 has the parameters a_i=1.0, b_i=0.0, and c_i=.18. Notice that high values of a_i and low values of c_i increase the amount of information available in an item. Also of interest is the fact that when c_i is non-zero, the information curve is skewed (reflecting the fact that guessing is effective) and thus has a greater effect on the lower end of the curve.

Figure 3
Information Curves



Research Plan

Selection of Test Items

Test items which best measure Word Knowledge (WK) and Arithmetic Reasoning (AR) abilities will be selected. A minimum of 180 WK and 180 AR items will be required. The criteria for pre-screening these items include:

1. Highest discriminatory power;
2. Difficulty values over the full ability range; and
3. Five alternative responses per item, if available.

The actual procedure involves "guesstimating" the item coefficient of guessing (c_i)=.20, and computing either the conventional point-biserial or biserial correlation between item-total test score and the proportion passing. Utilizing the charts provided in the literature (Urry, 1974, 1977), approximations of the item parameters are obtained. Jensema (1976) provides a computer program to generate approximations for item parameters using conventional item statistics.

Development of Parallel Tests

The test items will be divided into parallel test forms containing 60 test questions. The test forms will be generated by distributing the approximations for the item difficulties (b_i) widely and evenly throughout the ability range. This procedure will ensure a test information function which is reasonably flat over the entire ability range; this is important, since it provides a base for the manifestation of latent ability.

Failure to include a minimum of 60 test items or failure to measure all levels of ability evenly will provide an inadequate base to measure latent ability, resulting in inferior item parameter estimates. Monte carlo simulation (Gugel et al., 1976) has shown that 60 items is the minimum required for estimating reliable item parameters.

Administration of Test Booklets

Test booklets will be administered to a minimum of 2000 incoming Marines at the recruit depots. The examinees will be instructed to answer all 60 questions and to use as much time as necessary. The population at the recruit depots covers a full ability range of the target population (those personnel whose age makes them available for military service), thereby assuring that the parameters are properly located. Monte carlo simulation (Gugel et al., 1976) demonstrated that utilization of 2000 or more examinees will yield sufficiently reliable item paramter estimates.

If the test is treated as a speeded test rather than as a power test, two actions would ensue: (1) items at the end of the test would not be completed by the slower examinees; and (2) responses to items would be rushed, creating higher item discriminating powers (a_i), but lower item difficulties (b_i).

Estimating parameters with a population which is not representative of the target population will result in estimates of a_i and b_i which are algebraic transformations of the target population parameters. Using anchor items with both populations will enable determination of the differences in ability distribution. The algebraic transformation can be determined from the means and standard deviations of the two populations.

Estimation of Item Parameters

The item responses will be transferred from the answer sheets to computer input media. The item vectors will be processed via computer, utilizing the Urry (1976) item parameter estimation technique. This program cycles through two successive approximations of estimating the distribution of manifest ability: (1) corrected raw scores where the item being parameterized is omitted, and (2) using Bayesian modal ability estimates.

The output provides estimates of the parameters a_i, b_i, and c_i, using a smallest chi-square fit. After both cycles the approximations are then adjusted for the information of the group of items to increase their efficiency.

Adaptive Testing Item Bank Selection

The item bank requirements for efficient and accurate tailoring (Jensema, 1976a; Urry, 1974, 1977) are:

1. Item discriminatory powers (a_i) should be as high as possible, and at least .80.
2. Item difficulties (b_i) should be widely and evenly distributed.
3. Item coefficients of guessing (c_i) should be as low as possible, with .30 as a maximum.
4. Item bank should consist of a sufficient number of items.

Jensema (1976) has shown that when values of a_i become large and when value of c_i become small, greater tailoring efficiency occurs. This conclusion is consistent with the finding that greater information is available when an item has high a_i values and low c_i values. Where there are gaps in the distribution, failure to satisfy the requirement for a rectangular difficulty distribution will force the Bayesian algorithm to select an item which will yield less than optimal information about the examinee's ability.

Evaluation of Adaptive Testing Item Bank

An evaluation of the effectiveness of the adaptive testing item bank will be conducted through a simulation run of approximately 500 computer-generated "examinees." This will be accomplished by a monte carlo simulation of a normal (0,1) population for θ, using the estimated item parameters and the Bayesian ability estimation program for multiple ability banks. This procedure will allow an evaluation of the performance of the item banks in adaptive testing and will serve as a guide for assessing the future performance of testing live examinees.

Adaptive Test Administration

A computer-administered adaptive ability test will be administered to a sample of recruits at the Marine Corps Recruit Depot. This test will measure WK and AR abilities, using the Bayesian ability estimation program for multiple ability banks. The termination rule will be a pre-selected reliability, using the formula:

$$\sigma_\varepsilon = \sqrt{1-\rho^2_{\theta\hat{\theta}}} \qquad [4]$$

Using a standard error of estimate as a termination rule guarantees equiprecision of ability estimation for all examinees. Use of a maximum number of items as the termination rule will guarantee lower test reliability for upper ability examinees. The reason for this is that the Bayesian ability estimation process assumes a normal prior distribution of ability for all examinees. The effect of this prior distribution is shown in Figures 4 through 7.

Figure 4
$P'(\theta)(a_i=2.09,\ b_i=-.12,\ c_i=.17)$
and Posterior Distribution

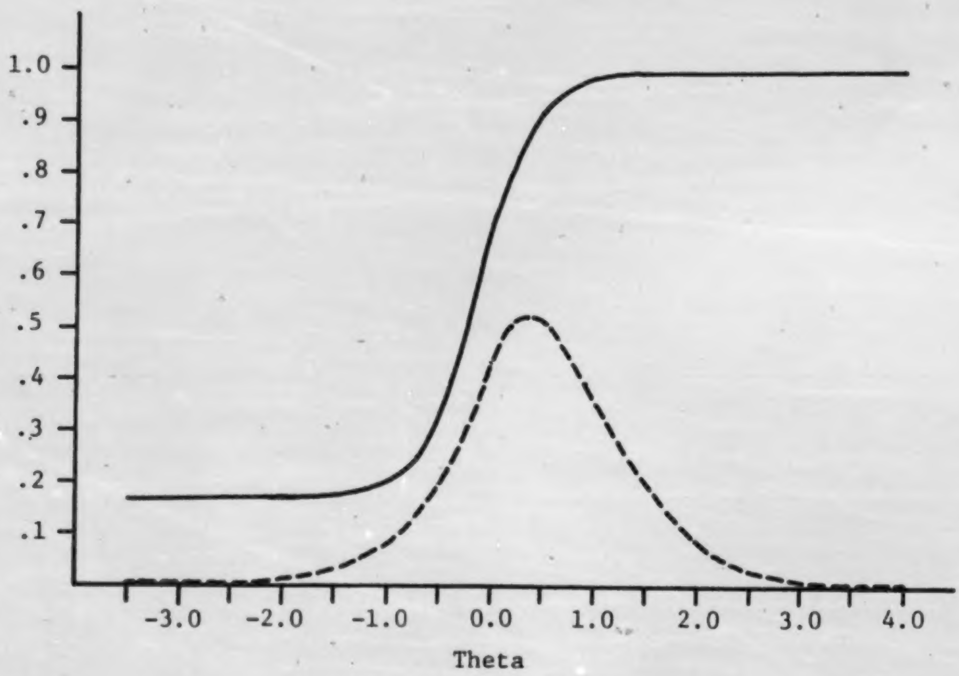
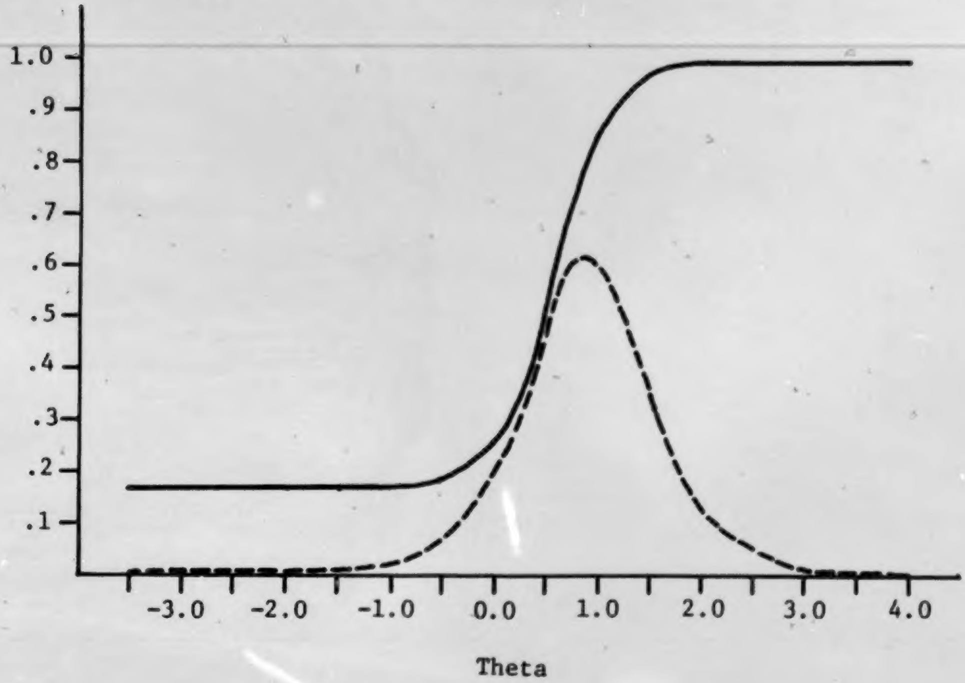


Figure 4 shows the posterior distribution after an examinee correctly responded to an item with a_i=2.09, b_i=-.12, and c_i=.17. At this point the ability estimate was .469, with standard error of .8568. Figure 5 shows the posterior distribution after the examinee had correctly responded to an item with a_i=2.25, b_i=.59, and c_i=.17. The ability estimate was then updated to θ=.929, with standard error = .7527. Figures 6 and 7 illustrate the greater capacity of an incorrect response to decrease the standard error of estimate. In Figure 6 the examinee responded incorrectly to an item with a_i=1.94, b_i=.78, and c_i=.13. The ability estimate was then .367 with standard error of .5549.

Figure 5
$P'(\theta)(a_i=2.25,\ b_i=.59,\ c_i=.17)$
and Posterior Distribution



Had the examinee responded correctly, as in Figure 7, the standard error would have decreased only to .6440. Since lower ability applicants will have more incorrect responses, their standard errors of estimate will decrease more rapidly. Those examinees who respond correctly to a greater proportion of items will require a longer test in order to maintain equiprecision of ability estimation.

Figure 6
$Q(\theta)(a_i=1.94,\ b_i=.78,\ c_i=.13)$
and Posterior Distribution

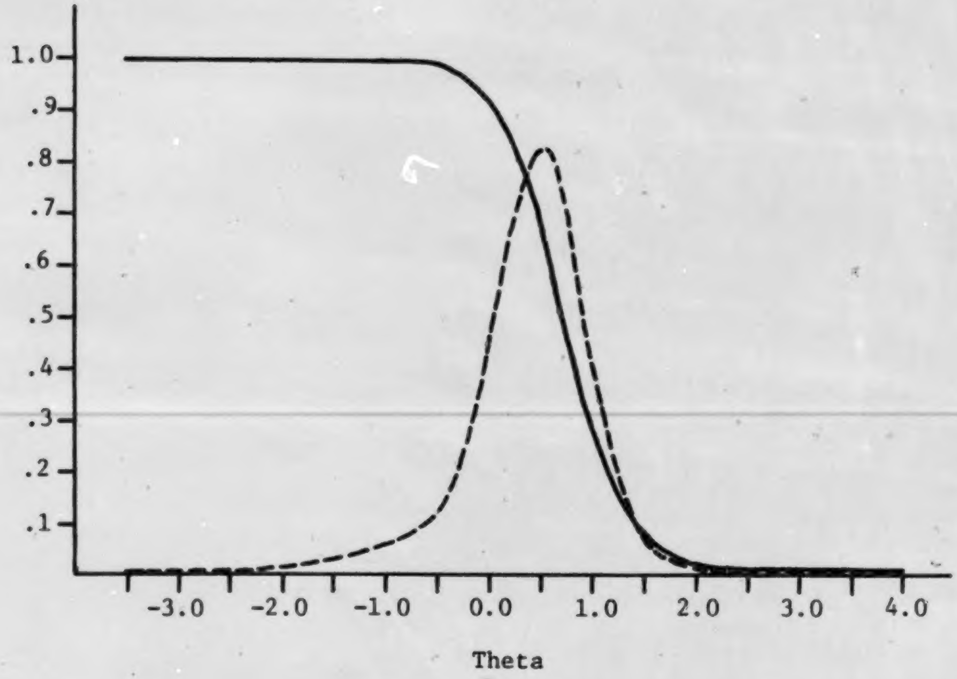
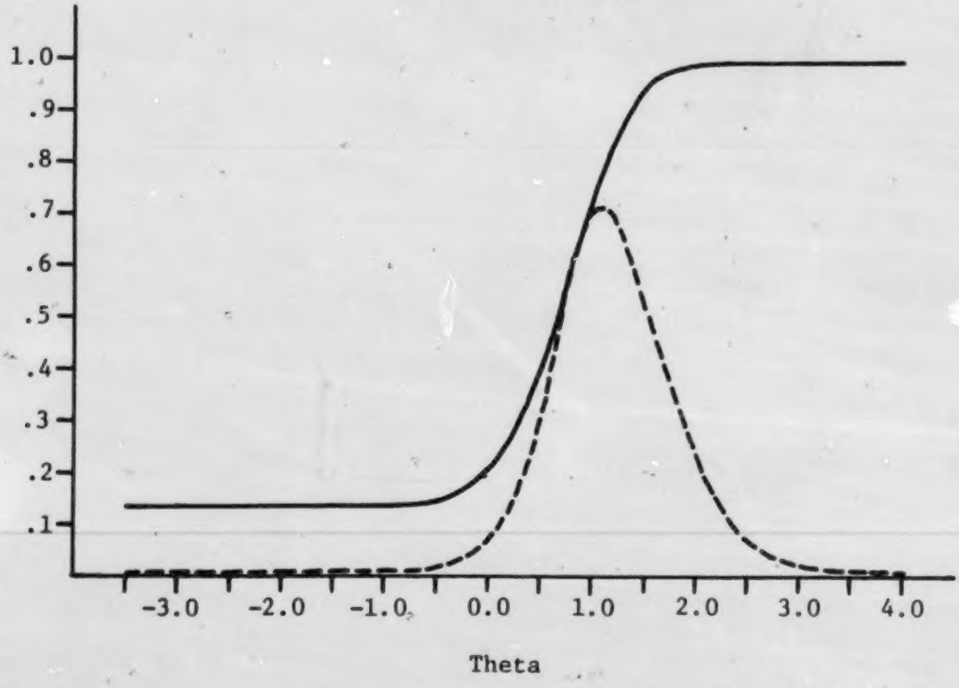


Figure 7
$P'(\theta)(a_i=1.94,\ b_i=.78,\ c_i=.13)$
and Posterior Distribution



Summary

This paper has demonstrated some of the errors which can be committed in instituting Bayesian adaptive testing. With an awareness and sensitivity to these and other potential problems, psychometricians can gain numerous benefits by using computerized adaptive testing.

References

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick (Eds.), Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

Gugel, J. F., Schmidt, F. J., & Urry, V. W. Effectiveness of the ancillary estimation procedure. Proceedings of the first conference on computerized adaptive testing (PS-75-6). Washington, DC: Personnel Research and Development Center, U. S. Civil Service Commission, 1976.

Jensema, C. J. Bayesian tailored testing and the influence of item bank characteristics. Computers and testing: Steps toward the inevitable conquest (PS-76-1). Washington, DC: Personnel Research and Development Center, U. S. Civil Service Commission, 1976. (a)

Jensema, C. J. A simple technique for estimating latent trait mental test parameters. Educational and Psychological Measurement, 1976, 36, 705-715. (b)

Owen, R. J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of the American Statistical Association, 1975, 70, 351-356.

Urry, V. W. A monte carlo investigation of logistic mental test models. (Doctoral dissertation, Purdue University, 1970). Dissertation Abstracts International, 1971, 31, 6319B. (University Microfilms No. 71-9475).

Urry, V. W. Approximations to item parameters of mental test models and their uses. Educational and Psychological Measurement, 1974, 34, 253-269.

Urry, V. W. A five-year quest: Is computerized adaptive testing feasible? Proceedings of the first conference on computerized adaptive testing (PS-75-6). Washington DC: Personnel Research and Development Center, U. S. Civil Service Commission, 1976.

Urry, V. W. Tailored testing: A successful application of latent trait theory. Journal of Educational Measurement, 1977, 14, 181-196.

IMPLEMENTATION OF TAILORED TESTING AT THE CIVIL SERVICE COMMISSION

RICHARD H. MCKILLIP
U.S. CIVIL SERVICE COMMISSION

The Civil Service Commission is moving carefully toward implementation of tailored testing. Apart from the prudence of gradual testing and installation of such a significant technical innovation, tailored testing presents significant budgetary and administrative problems that can be dealt with effectively only over time and with considerable study. We now have the technology to introduce tailored testing for cognitive abilities. What remains from a technical standpoint is to complete development of test question banks and to develop and implement additional improvements made possible by this new technology.

This paper discusses four stages in the implementation of tailored testing at the Civil Service Commission. In the first stage, the written test portion of the Professional and Administrative Career Examination (PACE) will be administered in the tailored version. Next, tailored testing will be extended to other examinations. Third, as latent trait theory is applied to all of the Civil Service Commission tests, a set of ability constructs will be identified that will be comprehensive enough to be useful in most employment settings. Finally, banks of questions testing these constructs will be developed.

Stage 1

In the first stage, a tailored version of the written test portion of the PACE will be administered. The feasibility study described by Segal (1977) calls for an operational tryout of tailored testing in one examining office in the fall of 1978. This experiment is expected to demonstrate the practicality of the method and to identify unanticipated operational problems. If the tryout is successful, full scale implementation of tailored testing for the PACE will be possible by the end of 1980.

To arrive at this point, several steps must be completed:

1. The necessary cathode ray tube units must be obtained and their appropriate distribution determined.
2. Necessary operational software must be developed. Software for presenting test questions and treating candidate responses already exists. What is needed is new software to provide information to candidates, to provide test scores in a form that is compatible with other information used in the examining process, and to

interface appropriately with the scoring and data processing system.

3. Details of test administration procedure must be planned.
4. Necessary announcements and publicity will need to be prepared.
5. A system for monitoring results will need to be developed to assure compatibility with the residual conventional testing, which for logistic reasons will probably continue for some time.

Stage 2

In the second stage of the implementation of computerized tailored testing, the capability for tailored testing will be expanded to other examinations. Additional question banks will be developed and existing banks expanded to reflect the different abilities required for successful performance of non-PACE jobs. Existing question banks have been parameterized on PACE applicant populations. New parameters will need to be developed for other applicant populations, because parameters will differ among groups of varying levels of ability.

Additional software will be developed and the computer network expanded to accomodate the growth of the system. In this stage some desirable modifications in hardware are expected to be identified, not only because experience with PACE administration will show up any deficiencies not previously anticipated, but also because tailored testing will provide a stimulus for development of new kinds of question types which will break the bonds of conventional multiple-choice paper-and-pencil questions. As tailored testing is integrated into the Commission's master system for processing examination data (SCORE, System of Comprehensive Operations for Recruiting and Examining) further software will be needed to achieve instant reporting of test results, automatic updating of eligible registers, speedy preparation of certificates of eligibles for particular job openings, and improved matching of candidates to particular job openings.

Stage 3

The third stage in implementing tailored testing at the Civil Service Commission will be concurrent with and somewhat independent of the second stage. Research will be carried out to identify a comprehensive set of constructs that will enable adequate coverage of most of the cognitive abilities that are relevant to employment testing. Work has already begun to calculate item statistics in accordance with latent trait theory. Eventually, there will be a finite set of constructs identified for each occupation through job analysis. Occupational categories will be established to include occupations with similar ability requirements. This contrasts with the typical classification approach which groups jobs and sets levels in accordance with similarity of duties, scope and complexity of work, responsibility, and other variables now directly related to ability to perform. This work is not specifically dependent on tailored testing but will provide an environment in which tailored testing can be maximally effective.

Stage 4

The fourth stage of implementation will be the construction of item banks for this comprehensive set of constructs and a system for relating item banks to jobs. At present examinations are developed for specific occupations or groups of occupations, which results in some duplication of effort. It also results in conflict for some job-seekers who conceivably might have to take several examinations in order to be sure they were considered for all the jobs which were of potential interest to them. Under the tailored system, an applicant would identify all occupations of interest and be examined for them in the most efficient way. The test then will be tailored both to the occupation and the individual applicant. It will even be possible to identify unique requirements of a specific job within an occupation and reconstitute an eligible register accordingly.

Further Implications of Computerized Testing

When these four stages are complete, tailored testing can be considered to be fully implemented. However, the administration of tailored tests is only the first of many possibilities for applications of a comprehensive system of tailored testing to Federal examining. The uses of employment and other tests have become subject to increasing scrutiny by the courts, by Federal and State enforcement agencies, by the psychological profession, and by the public. The need for comprehensive documentation of validity of tests has never been so intense. In many situations where the application of a particular test or item type is so routine that there is no question of the validity of the measure, critics of testing will nevertheless insist that comprehensive studies must be done. Additionally, and at least equally as important, the determination of the value of an examining program by means of utility analysis requires an estimate of the criterion-related validity of the examination.

Application of a tailored testing system will make available large amounts of data for validity studies as they are needed. Construct validation can be made automatic when the tailoring of item banks to occupations is complete. Criterion-related validity studies can be made more economically feasible by using the network of cathode ray tube terminals to obtain criterion measures from job incumbents and their supervisors without the expense of extensive field trips for data collection.

Applicant data other than test scores may be stored so as to expedite and improve hiring decisions. Evaluations of training and experience can be added to the system, and the pool of talent thus accumulated can be arrayed against existing or projected vacancies. The resulting improvement in job-person match will further improve hiring decisions, with the ultimate result of increased utility through improved performance.

The immediate availability of information will make it possible that an applicant, in one sitting at a video unit, could accomplish the following:

1. Take a test and be informed of performance,
2. Input experience, training and other information to be combined

with test score,
3. Be listed in appropriate eligible registers,
4. Receive a list of vacancies for which eligible,
5. Receive a statement of the probability of being hired, and
6. Receive further advice about job opportunities.

The immediate availability of information can provide hiring authorities with instant accessibility to the pool of qualified applicants. With the full information available about each applicant, hiring authorities will be able to make better decisions and more timely job offers, thus reducing the rate of declinations of offers and providing a better chance of obtaining the best available talent.

A fully developed system for computerized tailored testing of the abilities necessary to success in Federal jobs will bring about important improvements in every aspect of the process of selecting the most competent workers to fill them. It is for this reason that we look upon computerized tailored testing as the greatest advance in personnel testing since the group test was invented during World War I.

REFERENCE

Segal, H. Operational considerations in implementing tailored testing. Paper presented at the Computerized Adaptive Testing '77 Conference, Minneapolis, July, 1977.

OPERATIONAL CONSIDERATIONS IN IMPLEMENTING TAILORED TESTING

HAROLD SEGAL
U.S. CIVIL SERVICE COMMISSION

Based on the assumption that tailored testing is valid and feasible, the management of the U.S. Civil Service Commission has attempted to look at some of the practical problems involved with implementation of a large-scale tailored testing system and has established a task force that is interbureau and, to a great extent, interdisciplinary. The initial approach has been from the standpoint of applying marketing analysis techniques. It has, therefore, been necessary to look at the demographic characteristics of testing.

With regard to the Professional Administrative Career Examination (PACE), for example, the nationwide distribution of the locations where the test is given, the times of the year, and the concentration of test volume must be considered. The PACE Examination is administered in the 10 regional offices and in many of the 48 area offices of the U.S. Civil Service. About 90% of the applicants take the PACE Examination in approximately 12 or 14 locations throughout the country. This has implications for the potential computer configuration and the concentration of terminals needed. There will be a similar analysis of demographic requirements and test nature for the other examinations that are proposed for tailored testing before final feasibility judgements can be ascertained.

Organizational Problems

Organizational problems must also be addressed. The Civil Service Commission's Personnel Research and Development Center researches and develops tests. These tests are then channeled to the Bureau of Recruitment and Examination, which is composed of the operational personnel that operate the area offices and administer tests. The Bureau of Management Services controls the budget and is composed of management analysis and personnel specialists. Finally, the Bureau of Personnel Management Information Systems controls computer operations and manages the computer facilities; it is composed of analysts and programmers and is involved in all computer acquisitions.

A major problem is synchronizing all these organizational components and responsibilities within the total Civil Service Commission. If, in the near future, tailored testing is to become operational, solutions to managerial questions such as budget, logistics, hardware acquisitions, and software development are mandatory before operation of a major system can occur. However, while these problems are being resolved, there will be continued research and development for tailored testing on an experimental basis.

Budgetary Considerations

About 90,000 PACE tests were administered in the first half of 1976 and about 160,000 tests in the second half of the year. This volume of testing has strong practical implications concerning how terminals will be set up, how many terminals will be needed, and how the potential flow of applicants will be facilitated. Work is already in progress to determine costs and investment in the automation of a certain amount of the recruiting and examining processes, but much financial analysis will be necessary to determine how this would impact on the budgetary process and be carried out over a number of years. For a large-scale operation, lead time of approximately three years in the budgetary process is required. There are a number of problems in terms of computer acquisition, however, which may extend the lead time for a longer period. Therefore, if implementation is to occur in the early 1980s, such questions must be answered very soon.

Plans are now being made to define various alternative strategies concerning how to proceed with tailored testing with regard to hardware and software problems. Consultants are now being identified, and the scope of operations is being broadened to some extent. Although present thinking is in the direction of a large central processing operation, the cost of telecommunications could be very high; this remains to be seen. Exact costs need to be determined first. Other alternatives such as distributed processing and the use of microprocessors must be considered. It is this type of alternative analysis which needs to be addressed in terms of the costs and configurations that will be required.

The PACE Examination

Approximately a quarter of a million PACE exams are given in a year, with a seasonal distribution; and there is little likelihood of equalizing them to provide an even distribution of test administrations. There are perhaps 12 to 20 locations in the country where, on a given Saturday, the PACE exam is administered to between 500 and 1,000 people. If the paper-and-pencil exams were computerized, this would mean a potential of 250 to 500 terminals on an interactive basis, 8 hours a day, for those days (or periods or months) that the PACE exam would be administered. Aside from the costs, it is difficult to conceive of a centralized computer system which would be capable of handling the peak workloads that now exist.

Consideration of the cost is presently underway with a short-term feasibility study related to an operational test; however, a set of additional studies will be needed as each major decision point is approached. Some studies will probably be required to use queueing models, relating walk-ins, volume of tests, volume of applicants, and number of terminals. Initially, an operational test or prototype will be scheduled for 1980 or 1981. There will be an attempt to build into that prototype the kinds of questions and information necessary to subsequently build a model of what a computerized nationwide system would look like.

Other Considerations

From the cost standpoint, there are three broad categories which are traditional for large computerized system developments: (1) the development costs, (2) the implementation costs, and (3) the recurring operational costs. These categories of costs would be added to the current budget, since the paper-and-pencil system will continue and will overlap with the implementation of any new system.

The need for parallel operations must be considered in the cost analysis. The size of the agency is such that if the magnitude of the dollars required to begin the initial implementation is too large, it would be virtually impossible to obtain the level of funding necessary, aside from insufficient cost/benefit justification.

In addition, there must be some understanding of latent trait theory and the impact it has on testing, whether or not latent trait theory is applied to paper-and-pencil tests or computerized testing. Presumably, this will mean that in the near future, an operation such as the U.S. Civil Service may have to re-structure testing with regard to item banks and test environment, whether or not there is such a changeover. There is a way to dove-tail the costs of both types of testing: If those items are changed which relate to latent trait theory and there is a movement in the direction of tailored testing, a development cost is needed, whether the tests are computerized or not.

Conclusions

The problem is to have the idea described in a context that management or key decision-makers can understand: What is the magnitude of the proposal? What is the dollar amount involved? Can the logistics be worked out? Ultimately, there will be a point at which the management questions will need to be satisfied before computerized tailored testing can become an operational reality. These are very expensive questions which need to be carefully considered. A quantum leap has been taken by accepting the fact that tailored testing will probably work; the U.S. Civil Service will likely go ahead at this point in time with regard to what is required over the next few years to get the decisions through the entire governmental process for ultimate implementation. Presently, the managerial questions are a very weak link between research and application. If there is no link bridging the research and practical application research satisfies intellectual interests only, but from an operational standpoint there will be minimal success.

A LOW-COST TERMINAL USABLE FOR COMPUTERIZED ADAPTIVE TESTING

J. P. LAMOS AND B. K. WATERS
AIR FORCE HUMAN RESOURCES LABORATORY

The Technical Training Division of the Air Force Human Resources Laboratory (AFHRL/TT) at Lowry Air Force Base, Colorado, is currently involved in the development of a large computer-based instructional system, the Advanced Instructional System (AIS). The AIS is designed to handle a daily load of up to 2100 students within four Air Force technical training courses. As a computer-based system, the AIS provides primarily computer-managed instruction (CMI). In this context, computer-assisted instruction (CAI), including computerized adaptive testing, is one of the many pedagogical approaches which could be used; but due to present high terminal costs, the use of this approach is limited.

Cost, of course, is a major factor in the implementation of a computer-based instructional (CBI) system. Any CBI system, if it is to maintain cost-effectiveness, must be sensitive to the costs of courseware (instructional materials), software, and the capital investment in hardware. In terms of hardware, a CAI system is highly dependent on the cost of interactive terminals (e.g., Alpert & Bitzer, 1970). Computer costs per student for central site hardware and system software decline with an increase in students handled by the system (limited, of course, by the total capacity of the system). However, as students are added to a CAI system, total terminal costs increase dramatically, especially if each student is to be on the system for most of an instructional day.

The impact of terminal costs is reflected in the early cost projections for the PLATO IV system (Bitzer & Skaperdas, 1969). In attempting to partition total system costs between the various subsystems of PLATO, they estimated that central site hardware to support a 4,000-terminal system would be 4.5 million dollars for a cost per student contact hour of 8¢ attributable to the central site. The 4,000 terminals would cost 7.5 million dollars for a cost per student contact hour of 15¢ attributable to the terminals themselves. In this projection, terminal costs were almost twice central site costs. Actual experienced costs of PLATO IV hardware showed terminal costs to be almost five times as great as central computer facility costs (Alpert & Bitzer, 1970).

There are auxiliary support and instructional costs associated with CMI. In addition to the terminal hardware cost, there is a cost associated with the major means for communicating data through a management terminal--computer-readable forms. In the AIS these forms have been chemically treated to allow for immediate feedback of correct response to objective-type test questions. In one of the courses under the AIS, the cost of forms alone has been approximately $8,500 per year. Additionally, there may be operational costs associated with the storage, ordering, and handling of these forms. There are also the indirect costs resulting from the frustration that students experience from continuous darkening of small boxes on answer sheets to encode information.

Although the AIS is still under developement and costs have not stabilized, some indication of the relations between CMI terminal costs and the support costs associated with the terminals can be shown. An anticipated stabilized cost for one AIS management terminal in its present configuration is $18,000. The relation of capital investment costs to support costs over a five-year amortization in one of the four AIS courses is shown in Table 1.

Table 1
AIS Management Terminal and Forms Cost
per Student (4,000 Students/Year)

Terminal	
Three Terminals $54,000 Capital Cost	Five-Year Amortized Cost $2.70/Student
Forms	
One-Year Cost $8,500	Five-Year Operating Cost $2.13/Student

As can be seen, the cost of forms alone in a five-year period almost equals the cost of the terminals. Because form costs are operating costs, they will soon exceed the terminal investment. In the AIS, cost relationships are continually being analyzed. On the basis of such an evaluation, work was undertaken on the development of an electronic replacement for the paper forms--a low-cost responding device which, like a form, would be independent of the central computer system and yet easily interfaced with the management terminal of the AIS.

Analyzing the instructional environment of the AIS in relation to computer support indicated three major functional areas: (1) response handling, (2) information presentation, and (3) data collection. These three areas broadly cover the specific functional needs of a CAI terminal as reported by Martin, Stanford, Carlson, and Mann (1975). However, one should not opt for a "do everything" capability in one device unless one is ready to pay for capabilities which are not fully utilized or required. Thus, the primary philosophy behind the design of the AFHRL/TT device was to build in *only* those capabilities which were necessary in the AIS setting.

Design specifications were developed by an AFHRL/TT interdisciplinary team consisting of an electronics technician, a computer design engineer, a computer systems analyst, and educational/psychological researchers who tried to match specific instructional/testing requirements with computer hardware/software capabilities. It was seen that the type of responding device required in the AIS needed to have the interactive and dynamic response handling capabilities normally associated with CRT terminals and the low-cost and ease of use now associated with the electronic hand calculator. The device has been designated a "microterminal." The microterminal, its dynamic response handling capability combined with the cost-effective presentation of information provided by programmed tests or audiovisual materials, begins to fill the void of instructional requirements and computer usage which lie between the more traditional conceptions of CMI and CAI.

The focus of the microterminal is objective item testing, in the form of embedded or adjunct questions accompanying instruction and end of instruction achievement tests. There is a sizable body of literature represented by the work of R. C. Anderson, E. Z. Rothkopf, and L. T. Frase (to name a few major researchers) which has provided clear evidence for the instructional effectiveness of well-designed and appropriate questions. The decision was made to invest hardware development costs in the features which would have the greatest return on investment. Thus, the hardware design of the first prototype unit focused on a serial sequence of test items, with flexibility for control of feedback, response storage, and the retrieval of student response data. The initial prototype microterminal represented an interactive electronic test form.

Microterminal Development

The heart of the AFHRL/TT microterminal is a Motorola M6800 microprocessor along with one 1,024-bit random memory chip and six 4,096-bit programmable read-only-memory chips. This microcomputer has a capacity of up to 900 test item keys which are permanently stored and a capacity of up to 250 test items for which a student's response is temporarily stored. In the prototype model of the microterminal, there are four testing algorithms which control the responding situation. The device has the option of no feedback, immediate feedback, or delayed feedback as well as an answer-until-correct option. It collects test latency data in the form of running time to complete the test and could collect item latency data with software changes. The microterminal features a 16-key keyboard, 14 display lights to allow for the indication of established directive messages, 4 hexadecimal display units to allow for answer feedback, and a sequential form of key echoing for longer response items, such as the input of a Social Security Number.

The design of this prototype microterminal was thus limited to providing response management for the effectuation of data collection. A decision was also made to handle only objective-type responses. Without the need for either extensive presentation of information or constructed responses, the display features of the microterminal could be limited. Informational presentation is left to regular programmed materials. All design goals minimized cost while attempting to satisfy high probability instructional and testing requirements in the AIS. The cost design goal for the AIS microterminal development was $500 or less per unit. Figure 1 depicts the face of the prototype microterminal, and Figure 2 shows the overall design of the first prototype terminal.

Figure 1
Microterminal Faceplate

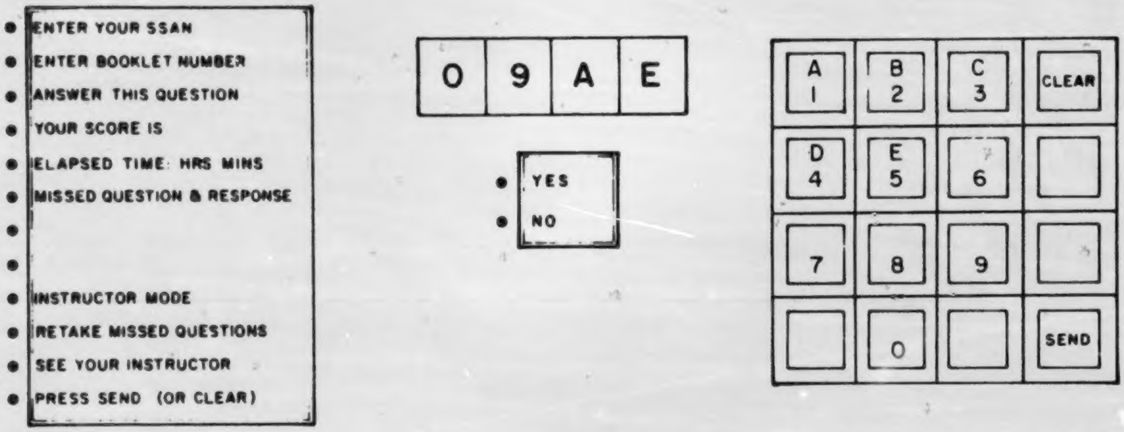

● ENTER YOUR SSAN
● ENTER BOOKLET NUMBER
● ANSWER THIS QUESTION
● YOUR SCORE IS
● ELAPSED TIME: HRS MINS
● MISSED QUESTION & RESPONSE
●
●
● INSTRUCTOR MODE
● RETAKE MISSED QUESTIONS
● SEE YOUR INSTRUCTOR
● PRESS SEND (OR CLEAR)
0 9 A E
● YES
● NO
A 1
B 2
C 3
CLEAR
D 4
E 5
6
7
8
9
0
SEND

Figure 2
Microterminal

A complete technical discussion of the microterminal hardware development is provided by Kirby and Gardner (1976).

Initial Microterminal Prototype Evaluation

A pilot study of a prototype microterminal was performed in order (1) to verify the usability and human factors of the design and (2) to attempt to gain a preliminary insight into potential instructional effects of the device. Results of the pilot study showed high user acceptance of the device, with no significant test grade differences between users and non-users. The study did indicate that students who used the microterminal accomplished lesson materials 30% faster than the control group. The time results of the pilot study are shown in Table 2. The difference in times between the groups may be due to a combination of factors such as time to fill out computer answer sheets, the effect of a computer directly monitoring the student's progress, increased attending behavior due to the direct tutorial effect (at least at the level of control over the response situations) of using the responding unit (i.e., less daydreaming), or simply a Hawthorne effect.

Table 2
Lesson Completion Times

	Control	Microterminal
N	12	12
Mean	168.08 minutes	117.75 minutes
S.D.	80.5	50.5

$(\overline{X}_c - \overline{X}_m) = 50.33$ minutes

$t = 1.85$, significant at $p = .05$ (one-tailed)

As part of the pilot study, an attitude questionnaire was administered to elicit student reactions to the device. The attitude questionnaire and student responses in percentages are shown in Table 3. Three items, in particular, seem important for systems design. Item 2 indicates that students preferred a device versus a form for recording their responses. Item 12 shows that a computer terminal device could be built to control the response feedback situation and to give clear directive-type information without an elaborate computer display. In other words, without having to pay for unused features such as a large display, the device satisfied an important part of the instructional situation; the information load gets carried by the programmed text or the test itself. Item 13 is of particular interest. Students subjectively reported that they "got more out of" the responder than if they used a sheet of paper to answer embedded questions. Normally, students are simply directed to record their answers on a sheet of paper, with feedback provided in the back of the programmed text. The microterminal has the advantage that the student must commit himself/herself to an answer before it is possible to receive feedback.

Table 3

Questionnaire Results of Students' Acceptance of the Microterminal (*N*=41)

1. What was your opinion of the responder? Good (86%) Bad (7%) Indifferent (7%)
2. Which do you like better? Responder (89%) Form (11%)

	Yes	No
3. Did you feel any type of pressure? If yes, why?	18%	82%
4. Did you feel nervous in using the responder? If yes, why?	20%	80%
5. Did you feel rushed? If yes, why?	7%	93%
6. Did you get frustrated? If yes, why?	18%	82%
7. Was someone present watching you use the responder?	47%	53%
8. Did this make you nervous in any way? If yes, in what way?	7%	93%
9. Do you feel uncomfortable about the computer grading your tests?	5%	95%
10. Is the device hard to use?	0%	100%
11. Could you understand the instructions in your lesson booklet? If not, how would you improve it?	93%	7%
12. Are the responder's red lights and their corresponding instructions easy to use?	100%	0%
13. Did you get more out of the programmed text questions using the responder rather than the usual way of answering on a sheet of paper or in you head?	88%	0%
14. Did you like the ability to continue responding until you received the right response for each question?	98%	2%

The results of any pilot study require more extensive confirmation; however, the present results adequately represent the viability of a particular hardware solution to an identified instructional need in a computer-based instructional system. The instructional implications of the use of a microterminal, as described in this report, can be significant. In the area of testing the use of different testing strategies, test item security, and automated entry of test results for item or response analysis by a larger system become administratively feasible or easier to implement. With more adequate control over the response-feedback element, directive control conveyed by algorithmically controlled indicator lights and changeable message cards, and properly formatted materials such as programmed texts, microfiches, or audiovisual presentations, a low-cost form of computer-assisted instruction becomes available.

Future Microterminal Development at AFHRL/TT

To further capitalize on the concept of low-cost CAI, AFHRL/TT has contracted for the development and evaluation of 10 newer prototype microterminals. The new microterminals will have several additional features, increasing the flexibility of response handling, data collection, and directive information for instructional and testing purposes. The new microterminal design still emphasizes the very limited display capabilities which resulted in significant cost reductions.

A major new feature is an insertable, self-contained memory module. The initial prototype was designed to be completely portable so that the self-contained student response data could be transferred to central site storage by having the student carry the microterminal to a CMI management terminal and plug it in for transfer of data and refreshment of the microterminal's RAM. Further examination of potential microterminal utilization revealed that although total size of the unit should be kept small to allow ease of use in student carrels, the device itself need not be carried every time transfer of data was required.

As in the original design, hardwiring of terminals was not desirable; there was a need to avoid the complexity of real-time communications between central site (i.e., polling of terminals, buffering of data) and a large number of terminals. The solution was to externalize that portion of memory which temporarily stores the students' responses, making *it* self-contained and portable. The additional benefit of this design is that the programmable hardware of the microterminal can be stabilized while the external memory unit provides a potentially expandable capability to support additional requirements. This becomes important in the context of adaptive testing,where some algorithms, such as maximum likelihood estimation and Bayesian scoring procedures, have relatively large processing and memory requirements.

Since the microterminal has limited display features, test or instructional information must be presented through more conventional media. The software of the initial microterminal was designed for serial presentation of test items. The new prototype will accommodate flexible or adaptive item presentation sequences. There are several ways in which either test or instructional materials can be presented to a student. One of the most flexible and best suited to adaptive presentation is microfiche.

The typical microfiche has 240 images, allowing for the presentation of both verbal and graphic materials. The row by column format of microfiche allows easy accessability of a particular image. The one significant problem associated with microfiche is the controlled access of an image. In the adaptive testing situation, it is most important to insure that a student look at the test item to which he/she was directed by the microterminal. To this end, the new microterminal will have an input/output port which will allow "communication" with a special electronic grid attached to a conventional microfiche projector. After a student has located the microfiche, the electronic grid will monitor and verify this location with the directed item given by the microterminal program. The microterminal and microfiche projector in combination provide a low-cost control and presentation "terminal" system.

Projected Implementation

The use of the Air Force's microterminal and a microfiche projector for adaptive testing may be as follows: A student would report to the testing center. Upon reporting to the center monitor, he/she would receive a memory module containing refreshed memory storage for responses, adaptive test parameter information, and a test microfiche. The student would be seated at a test carrel, insert the memory module into the microterminal, and insert and index the microfiche into the projector. An indicator light would illuminate beside a standard message requesting the student to enter his/her identification information. If this were accomplished correctly, another light would come on beside a message asking the student to enter a test number; the test number activates the adaptive testing algorithm. For example, the microterminal may indicate that the student should locate and answer Item 30. Should the student incorrectly move the microfiche to the location of Item 31 in attempting to answer the test item, his/her response would not be accepted. Instead, a message light would come on, indicating that the student was at the wrong location. Correctly readjusting the microfiche location, the student would note that the correct location light was on; and the microterminal would accept the answer. Depending on feedback condition, the student may or may not be told if the response is correct. The student then would continue, complete the test, and the microterminal would indicate final score.

Taking the memory module, the student would report to the test center monitor and place the memory module in a management terminal connected to the central site computer. The student's total test data would then be transferred to permanent computer records. Immediate test analysis and results may be requested and received through the management terminal by the test monitor, if so desired.

In addition to test center activities, microterminals may be remotely located with memory modules delivered or mailed for later processing and reuse.

Recommendations to Agencies Developing Adaptive Testing Terminals

It is recommended that other agencies or researchers interested in the development of microterminal-like devices for their own uses consider the following suggestions:

1. Begin with an extremely clear delineation of the requirements of your specific application.

2. Do not try to adapt an existing system design to another situation, unless the design requirements are compatible. For example, an agency desiring a device for adaptively testing student ability may not need a device like the AFHRL/TT microterminal, which was designed for an instructional environment; and paying for capabilities that are not utilized is unsound.

3. Use an interdisciplinary team to specify the design of the device. Infeasible goals for the device in terms of cost, size, or capability

would quickly be recognized by the computer systems designer or the financial manager.

4. Finally, keep abreast of the rapidly changing microprocessor technology. Designs are practically outmoded before they are even built. Capabilities are expanding geometrically while costs are decreasing arithmetically, yielding large increases in capability per dollar. However, it should be realized that at some point a design commitment must be made, irrespective of future technological advances.

Research and development in terminal design at AFHRL have demonstrated the feasibility of developing a low-cost testing microterminal for use in AIS instruction. Combined with other developments in the psychometrics of adaptive testing we appear to be at the edge of what Green (1970) predicted would be "the inevitable computer conquest of testing."

References

Alpert, D., & Bitzer, D. L. Advances in computer-based education. Science, 1970, 167, 1582-1590.

Bitzer, D., & Skaperdas, D. The design of an economically viable large-scale computer-based education system (CERL Report X-5). Champaign-Urbana, IL: University of Illinois, Computer-Based Education Laboratory, 1969.

Green, B. G., Jr. Comments on tailored testing. In W. H. Holtzman (Ed.), Computer-assisted instruction, testing, and guidance. New York: Harper & Row, 1970.

Kirby, P. J., & Gardner, E. M. Microcomputer controlled, interactive testing terminal development (AFHRL-TR-76-66). Lowry Air Force Base, CO: Air Force Human Resources Laboratory, Technical Training Division, 1976.

Martin, T. H., Stanford, M. D., Carlson, F. R., & Mann, W. C. A policy assessment of priorities and function needs for the military computer-assisted instruction terminal (ISI/RR-75-44). Marina del Rey, CA: University of Southern California, Information Sciences Institute, December 1975.

Editor's Note

This paper was not formally presented at the Conference. However, since the microterminal was demonstrated during the Conference, this paper was included to properly document this important development.

SESSION 6

ACHIEVEMENT/PERFORMANCE TESTING VIEWED AS A CLASSIFICATION PROBLEM

APPLICATIONS OF SEQUENTIAL TESTING PROCEDURES TO PERFORMANCE TESTING	KENNETH I. EPSTEIN AND CLARAMAE S. KNERR ARMY RESEARCH INSTITUTE
ADAPTIVE BRANCHING IN A MULTI-CONTENT ACHIEVEMENT TEST	ROGER PENNELL AIR FORCE HUMAN RESOURCES LABORATORY
ADAPTIVE TESTING APPLIED TO HIERARCHICALLY STRUCTURED OBJECTIVES-BASED PROGRAMS	RONALD K. HAMBLETON AND DANIEL R. EIGNOR UNIVERSITY OF MASSACHUSETTS
MULTI-CONTENT ADAPTIVE MEASUREMENT OF ACHIEVEMENT	DAVID J. WEISS AND JOEL M. BROWN UNIVERSITY OF MINNESOTA
DISCUSSION	RICHARD L. FERGUSON AMERICAN COLLEGE TESTING PROGRAM

SESSION 6: ABSTRACTS

APPLICATIONS OF SEQUENTIAL TESTING PROCEDURES TO PERFORMANCE TESTING

KENNETH I. EPSTEIN AND CLARAMAE S. KNERR

Wald's acceptance-sampling procedures and mathematics are expressed in terms of (1) classifying students with respect to their mastery of instructional objectives and (2) classifying instructional materials with respect to their instructional effectiveness. Specification of necessary parameters is discussed in the context of practical applications, and sampling plans and decision rules are developed. Techniques for determining the expected number of observations necessary for making decisions and for constructing operating-characteristic curves are presented. The procedures, which were applied to the evaluation of Army audio-visual packages and to the assessment of trainees on a variety of performance tasks, are discussed in terms of (1) the practical impact of violations of the assumptions underlying the procedures and (2) the validity of the classification decisions. Both existing and possible future uses of the procedures with computer assistance are described.

ADAPTIVE BRANCHING IN MULTI-CONTENT ACHIEVEMENT TESTS

ROGER PENNELL

Two simulation studies were performed in order to explore potential test time savings derivable from alternate flexilevel strategies. Study I explored the potential of entering students at items other than the median item. Flexilevel testing was simulated on 186 paper-and-pencil protocols obtained from an Air Force test and in data from 100 students who took a standard flexilevel test. Item savings were approximately 50%. Study II used 133 test protocols obtained from students tested on a battery comprised of five subscales of decreasing difficulty and items within scales of decreasing difficulty. More than 50% of the test items were saved in each block, and the resulting flexilevel score correctly classified 100% of the students in Block II and 94% in Block IV. The results of the two studies suggest that using alternative flexilevel strategies, considerable test savings are potentially available at relatively little cost.

MULTI-CONTENT ADAPTIVE MEASUREMENT OF ACHIEVEMENT

DAVID J. WEISS AND JOEL M. BROWN

An adaptive testing strategy combining adaptive item selection both within and between the subtests in a multiple-subtest battery is described for use with achievement tests which cover multiple content areas. A real-data simulation compared the results from adaptive testing with those from conventional testing in terms of test information and test length, using data from 365 fire control technicians on a 232-item achievement test battery with 12 subtests. Correlations between subtest scores from adaptive and conventional testing were .90 or higher for 11 of the 12 content areas; and an information analysis resulted in essentially identical information curves for all 12 subtests for the adaptive and conventional strategies. The number of items administered with adaptive testing was, on the average, 50% of that required with conventional testing.

ADAPTIVE TESTING APPLIED TO HIERARCHICALLY STRUCTURED OBJECTIVES-BASED CURRICULA

RONALD K. HAMBLETON AND DANIEL R. EIGNOR

The purposes of this paper are: (1) to introduce the nature of objectives-based curricula, criterion-referenced tests which are used in these programs, and the kinds of testing problems that are encountered; (2) to review methods for validating criterion-referenced test items and for establishing the hierarchical structure of a set of instructional objectives; and (3) to introduce several adaptive testing strategies and results with a hierarchically structured set of objectives. Computer simulation results are reported, showing that it is possible to obtain a reduction of more than 50% in testing time using adaptive testing compared to a conventional testing procedure, without any loss in instructional decision-making accuracy.

APPLICATIONS OF SEQUENTIAL TESTING PROCEDURES TO PERFORMANCE TESTING

KENNETH I. EPSTEIN
CLARAMAE S. KNERR
U.S. ARMY RESEARCH INSTITUTE
FOR THE BEHAVIORAL AND SOCIAL SCIENCES

Methods for accurately and efficiently making classification decisions are becoming more necessary in applications of educational and psychological testing. The systematic design of instruction requires that data indicating the degree of effectiveness of instructional materials be collected and used to identify that which is acceptable and that which needs improvement. Criterion-referenced testing implies that students will be classified on the basis of whether or not they have mastered specified instructional objectives. Performance testing usually refers to a special case of criterion-referenced testing that applies to training programs emphasizing job-related skills. Trainees are classified on the basis of performance test scores as sufficiently skilled or inadequately trained.

Typically, when instructional effectiveness is evaluated, the classification decision is dependent upon whether or not some criterion proportion of students successfully completes a program of instruction. For the criterion-referenced or performance testing case, classification depends upon whether or not the criterion proportion of test items is passed. This paper is concerned with two problems associated with these procedures: (1) the number of students or items tested and (2) the specification of misclassification error rates. Before addressing these problems directly, it will be necessary to clarify the testing problem and to state several important assumptions concerning the nature of the data.

The Classification Problem in Testing

Baker and Alkin (1973) have suggested that one of the critical factors in judging instructional effectiveness is the extent to which learners master the objectives. This can be taken one step further to conceptualize instructional effectiveness in terms of the extent to which *any* student in the target population is likely to master the objectives, given the opportunity. This is exactly what is implied in the 80/80 criterion often applied to instructional development efforts; the instruction is considered effective if at least 80% of the students who begin the instruction complete at least 80% of the objectives on the first attempt.

Criteria of Instructional Effectiveness

Let it be assumed, for the moment, that the 80/80 criterion is reasonable for a particular instructional development effort. That is, although all of the students who begin the instruction may not successfully complete 80% of the objectives on the first attempt, if 80% of them do, it will be satisfactory. Let it also be assumed that valid and reliable measures of the objectives are available. How is it determined whether or not the instruction is acceptable in its present state? The obvious answer is to find subjects who are members of the population for which the instruction is intended, to let them try the instructional materials, and to find out how well they perform on a test of the objectives.

Under the 80% criterion rule, the instruction will be considered effective if at least 80% of the students in the target population accomplish the objective. This criterion may also be interpreted as the probability that any randomly chosen student will accomplish the objective, that is, .80. In other words, the performance of a randomly chosen student may be considered a Bernoulli variable, with the probability of success equal to instructional effectiveness. Either an acceptance or rejection decision can be made when sufficient evidence to draw inferences about the instructional effectiveness has been gathered.

Sample Size

This leads directly to the question of the size of the tryout sample of students. The data gathered during a tryout of the instruction which is designed to meet this criterion is used to draw inferences about the effectiveness of the instruction for the total target population. Clearly, the larger the sample of students in the tryout group, the better the estimate of instructional effectiveness for the total group will be.

An indication of the precision with which population parameters are estimated by sample data is given by confidence limits. For example, assume that the tryout sample for one objective consists of five students, four of whom accomplish the objective. The effectiveness of the instruction, in terms of the proportion of students who accomplished the objective, is 80%. However, the 95% confidence limits for a proportion based on four correct answers in five trials are .343 and .990. These confidence limits assume that a random sample is drawn from an infinitely large population. Since most instructional development efforts involve materials which will be useful for a large number of students and since students for a tryout should be randomly sampled from the target population, this assumption seems reasonable.

The relatively widely separated values of the confidence limits imply that extreme caution should be taken in drawing any inferences about the instructional effectiveness for the total population from a tryout sample of five. Unfortunately, increasing the sample size, while staying within the bounds of practical constraints, is not very helpful. For example, the 95% confidence limits for a proportion based on observing 8 correct in 10 trials are .397 and .963; for 16 correct in 20 trials, .589 and .929; for 24 correct in 30 trials, .636 and .909; for 40 correct in 50 trials, approximately .67 and .90; and for 80 correct in 100 trials, approximately .71 and .88. Novick and Lewis (1974) discuss the sample size problem in terms of errors in decision making from a criterion-referenced point of view. Their conclusions are equally discouraging for small samples of data.

Misclassification

The misclassification question is also difficult, partly because it is rarely addressed or understood in practical applications of instructional materials development. The first problem is choosing a criterion proportion of students in the target population for whom the instruction should be effective. Ideally, all students should be reached; but this is an unrealistic goal. Individual student differences, costs involved in designing instructional materials, and limitations on instructional time and resources all mitigate against achieving this ideal. Consumers and developers of instructional materials must begin to consider these factors and set reasonable goals for themselves based both on their specific needs and on the student populations. While the sequential testing procedure to be described in this paper does not offer a method for determining the criterion, it does demand that a criterion be specified. Perhaps by requiring such specificity, discussions necessary to arrive at meaningful criteria will transpire.

A second requirement of the procedure is that the decision maker realize that since he/she is only able to sample part of the target population, occasionally incorrect decisions will be made. Thus, very good instruction will sometimes be rejected,based on tryout results--a false negative decision--and poor instruction will sometimes be accepted--a false positive decision. The costs of such misclassifications must be considered in deriving a decision-making procedure. If misclassifications are considered extremely costly, then a larger tryout sample must be used. If the costs of false positive and false negative decisions are not equal, then adjustments must be made in the decision rule to insure that the likelihood of the more costly incorrect decisions is kept small. As in the case of the criterion associated with acceptable instruction, the sequential testing procedure does not solve the misclassification rate problem. The procedure does, however, force decision makers to come to grips with the problem.

Criterion-Referenced and Performance Testing

The same types of problems faced by instructional materials designers are encoutered by criterion-referenced and performance test developers. The test is typically a relatively small sample of items or problem situations drawn from a large pool defined by the objectives. Time and resource constraints require that the number of items be kept small. Mastery of the objective is stated as a criterion proportion of correct responses that would be obtained if all items in the pool could be administered. A decision rule is required that classifies students as masters or non-masters of an objective, based on their responses to the sample of items included on the test.

The test designer must answer questions analogous to those of the instructional designer: (1) What should the criterion proportion correct be?

(2) How many items should be included on the test? (3) What are the relative costs of false positive and false negative decisions concerning student ability? (4) How can they be quantified? and (5) How can these factors be incorporated in a usable decision rule?

Sequential Analysis in Instructional Design and Criterion-Referenced Testing

The following assumptions apply to the data collected both for instructional materials evaluation and student evaluation. The data are a series of dichotomous pass-fail decisions. In each case, the sample proportions of pass-fail decisions represent estimates of those proportions that would be obtained if all students in the target population or all items in an objective's domain could be tested. Each student or item is considered an independent random sample from the population and has associated with it a probability of passing equal to the population proportion of passing decisions. The particular sequence of students tested or items presented does not represent a systematic order effect. Each individual student or item pass-fail decision is unambiguous.

Sequential analysis was developed as an alternative to traditional statistical hypothesis testing by Wald in the early 1940s. The theory and techniques are described in Wald's Sequential Analysis[1] This paper will be concerned only with the application of the general theory to testing the mean of a binomial distribution (Wald, 1973, pp. 88-105); the mathematics involved will be described and the necessary parameters in terms of instructional effectiveness and criterion-referenced testing will be defined. This will be followed by a computed example and the results of three applications of the procedure.

Procedures of Sequential Analysis

Wald's context for describing the procedure was acceptance inspection of a lot consisting of a large number of manufactured products. The problem was to determine whether or not the proportion of defective items exceeded some predetermined limit based on the inspection of a relatively small sample of items. Similarly, the instructional design and criterion-referenced testing problems are to determine whether or not the criterion proportions required for acceptance were exceeded by (1) the proportion of students for whom the instruction was ineffective and (2) the proportion of items answered incorrectly on a criterion-referenced test.

Accept/Reject decisions. Let p equal the population proportion of students for whom instruction is ineffective or the proportion of items an individual student would answer incorrectly, given all the items in the domain. Let p_c equal the corresponding criterion proportions required for acceptance. If $p<p_c$, the correct decision is to accept. If $p>p_c$, the correct decision is to reject. Since decisions will be based on sample rather than population data, misclassification errors must be considered. When $p<p_c$, the preference to accept increases with decreasing values of p. When $p>p_c$, the preference to reject increases with increasing values of p. However, since errors will occur, it should be possible to define some p_o slightly below p_c where an incorrect rejection decision has little practical consequence. Similarly, some p_1 slightly above p_c can be chosen where an incorrect acceptance decision is not serious. The region $p_o<p_c<p_1$ is known as an indifference region. If the population proportion, p, falls within the indifference region, the practical consequences of an incorrect decision are negligible.

---

[1] Page references in this paper refer to the unabridged and unaltered republication of the 1947 work by Dover Publications, Inc., 1973.

Once the limits of the indifference region have been specified, it becomes reasonable to choose values for the risks of incorrect decisions. The probability of a rejection decision, when $p\leq p_o$, should not exceed some small value α; and the probability of an acceptance decision, when $p\geq p_1$, should not exceed some small value β. In other words, α is the acceptable risk of committing a false negative error and β is the acceptable risk of committing a false positive error.

Given values for the following five parameters, a sequential sampling plan can be specified:

1. p_c, the population proportion or probability of failures or incorrect responses that defines an unacceptable product or student.
2. $p_o < p_c$, a lower limit proportion or probability of failures or incorrect responses below which false negative errors are critical;
3. $p_1 > p_c$, an upper limit proportion or probability of failures or incorrect responses above which false positive errors are critical;
4. α, the acceptable risk of committing a false negative error; and
5. β, the acceptable risk of committing a false positive error.

The choice of the parameter values is not a statistical problem. Rather, the values must be chosen on the basis of the practical considerations and requirements of each particular test.

Sequential sampling. The sequential sampling procedure takes advantage of the need for relatively little data to identify a very good or a very poor product or student and the requirement of extensive observations only for marginal cases. For each observation, the probabilities of the observation and all preceding observations, given $p=p_o$ and $p=p_1$, are calculated and the log of their ratio obtained. An accept, reject, or continue sampling decision is made based on the value of the log probability ratio. Wald has shown that an accept or reject decision will eventually occur with probability 1.0 and that fewer observations are required, on the average, for given values of p_o, p_1, α, and β using the sequential sampling procedure than are required by traditional hypothesis testing procedures.

Let m equal the number of misses or failures in the first n observations. Given $p=p_o$, the probability of observing such a sample is

$$p_{om} = p_o^{\,m}(1-p_o)^{n-m} \quad . \qquad [1]$$

For $p=p_1$, the probability is

$$p_{1m} = p_1^{\,m}(1-p_1)^{n-m} \quad . \qquad [2]$$

The log of the ratio is now computed

$$\log \frac{p_{1m}}{p_{om}} = \log \frac{p_1^{\,m}(1-p_1)^{n-m}}{p_o^{\,m}(1-p_o)^{n-m}} = m \log \frac{p_1}{p_o} + (n-m) \log \frac{(1-p_1)}{(1-p_o)} \quad . \qquad [3]$$

After each observation, the following inequality is evaluated:

$$\log \beta < \log \frac{p_{1m}}{p_{om}} < \log A \quad . \qquad [4]$$

If $\log \frac{p_{1m}}{p_{om}} > \log A$, then the test terminates with a rejection decision.

If $\log \frac{p_{1m}}{p_{om}} < \log B$, then the test terminates with an acceptance decision.

Otherwise, another observation is made and the decision rule is applied again.

The values of A and B are functions of α and β. Wald (1973, pp. 41-48) discusses the relationships between A, B, α, and β and shows that very close approximations to the exact values required for hypothesis testing are obtained by setting

$$A = [(1-\beta)/\alpha] \qquad [5]$$

and

$$B = [\beta/(1-\alpha)] \quad . \qquad [6]$$

In fact, $[(1-\beta)/\alpha]$ and $[\beta/(1-\alpha)]$ are upper and lower limits for the exact values of A and B, respectively, which implies that the use of the approximations will provide a slightly conservative test.

Replacing A and B with their approximations and performing several algebraic steps leads to a convenient form of the decision-making inequality:

$$\log[\beta/(1-\alpha)] < m \log \frac{p_1}{p_o} + (n-m) \log \frac{(1-p_1)}{(1-p_o)} < \log\,[(1-\beta)/\alpha] \qquad [7]$$

$$\log[\beta/(1-\alpha)] < m \left(\log \frac{p_1}{p_o} - \log \frac{(1-p_1)}{(1-p_o)}\right) + n \log \frac{(1-p_1)}{(1-p_o)} < \log \frac{1-\beta}{\alpha} \qquad [8]$$

$$\frac{\log \frac{\beta}{1-\alpha} - n \log \frac{(1-p_1)}{(1-p_o)}}{\log \frac{p_1}{p_o} - \log \frac{(1-p_1)}{(1-p_o)}} < m < \frac{\log \frac{1-\beta}{\alpha} - n \log \frac{(1-p_1)}{(1-p_o)}}{\log \frac{p_1}{p_o} - \log \frac{(1-p_1)}{(1-p_o)}} \quad . \qquad [9]$$

Using this form of the equation, the number of misses (m) is compared to the extremes of the inequality. If m is less than or equal to the left hand extreme, the test terminates with an acceptance decision. If m is greater than or equal to the right hand extreme, the test terminates with a rejection decision; otherwise, sampling continues. An interactive computer scoring scheme would simplify this process.

Graphical procedures. If a computer is not available, a graphic form of the procedure can be used. The abscissa represents the number of observations; the ordinate represents the number of misses. The extremes of the inequality represent two parallel lines on the graph with common slope:

$$\frac{-\log \frac{(1-p_1)}{(1-p_o)}}{\log \frac{p_1}{p_o} - \log \frac{(1-p_1)}{(1-p_o)}} = s \qquad [10]$$

and intercepts

$$\frac{\log \frac{\beta}{1-\alpha}}{\log \frac{p_1}{p_o} - \log \frac{(1-p_1)}{(1-p_o)}} = I_a \qquad [11]$$

and

$$\frac{\log \frac{1-\beta}{\alpha}}{\log \frac{p_1}{p_o} - \log \frac{(1-p_1)}{(1-p_o)}} = I_r \qquad [12]$$

for the left and right extremes respectively. Each datum is plotted as it is observed. As soon as the graph of the observations crosses the upper line, the test terminates with a rejection decision. As soon as it crosses the lower line, the test terminates with an acceptance decision; otherwise, the sampling continues. Using s, I_r, and I_a to represent the common slope, the intercept of the rejection line, and the intercept of the acceptance line, the decision rule can be simply stated:

1. Reject if $m \geq I_r + ns$;
2. Accept if $m \leq I_a + ns$;

3. Continue sampling if $I_a + ns < m < I_r + ns$.

Several other useful functions are also easy to calculate. The minimum number of observations--all misses--required for a rejection decision is the next higher integer greater than $[I_r/(1-s)]$. The minimum number of observations--all passes--required for an acceptance decision is the next higher integer greater than $(I_a/-s)$. The operating characteristic (OC) function of the test, $L(p)$, is the probability of an acceptance decision for p, the population proportion of misses. It is usually only necessary to calculate five values of $L(p)$ to graph the OC function, since the general shape of the curve is similar for any sequential sampling plan.

Five convenient values are

1. $L(o) = 1$; for $p=o$ the test will always terminate eventually with an acceptance decision
2. $L(1) = 0$; for $p=1$ the test will always terminate eventually with a rejection decision
3. $L(p_o) = 1-\alpha$, by definition
4. $L(p_1) = \beta$, by definition

5. $$L(p=s) = \frac{\log \frac{1-\beta}{\alpha}}{\log \frac{1-\beta}{\alpha} + \left|\log \frac{\beta}{1-\alpha}\right|} \qquad [13]$$

$$= \frac{I_r}{I_r + |I_a|}$$ (Wald, 1973, p. 95).

Wald (1973, pp. 96-98) provides a general equation for any value of $L(p)$:

$$L(p) = \frac{\left(\frac{1-\beta}{\alpha}\right)^h - 1}{\left(\frac{1-\beta}{\alpha}\right)^h - \left(\frac{\beta}{1-\alpha}\right)^h}, \qquad [14]$$

where h can take any value between $\pm \infty$ and

$$p = \frac{1-\left(\frac{1-p_1}{1-p_o}\right)^h}{\left(\frac{p_1}{p_o}\right)^h - \left(\frac{1-p_1}{1-p_o}\right)^h}. \qquad [15]$$

Finally, the expected number of observations, $E_p(n)$, necessary to reach a decision can be calculated as a function of p (Wald, 1973, pp. 99-101). The general equation is

$$E_p(n) = \frac{L(p) \log \frac{\beta}{1-\alpha} + [1-L(p)] \log \frac{1-\beta}{\alpha}}{p \log \frac{p_1}{p_o} + (1-p) \log \frac{1-p_1}{1-p_o}} \qquad [16]$$

Special formulas for the values of p used to calculate the OC function are

$$E_{p=o}\ (n) = \frac{\log \frac{\beta}{1-\alpha}}{\log \frac{1-p_1}{1-p_o}} \qquad [17]$$

$$E_{p=1}\ (n) = \frac{\log \frac{1-\beta}{\alpha}}{\log \frac{p_1}{p_o}} \qquad [18]$$

$$E_{p=p_o}(n) = \frac{(1-\alpha) \log \frac{\beta}{1-\alpha} + \alpha \log \frac{1-\beta}{\alpha}}{p_o \log \frac{p_1}{p_o} + (1-p_o) \log \left(\frac{1-p_1}{1-p_o}\right)} \qquad [19]$$

$$E_{p=p_1}(n) = \frac{\beta \log \frac{\beta}{1-\alpha} + (1-\beta) \log \frac{1-\beta}{\alpha}}{p_1 \log \frac{p_1}{p_o} + (1-p_1) \log \left(\frac{1-p_1}{1-p_o}\right)} \qquad [20]$$

$$E_{p=s}\ (n) = \frac{-\log \frac{\beta}{1-\alpha} \log \frac{1-\beta}{\alpha}}{\log \frac{p_1}{p_o} \log \left(\frac{1-p_o}{1-p_1}\right)} \qquad [21]$$

where $E_{p=s}\ (n)$ is or very nearly approximates the maximum value of the function.

Numerical Example

The three applications of the procedure (discussions of which follow) used the same values for the necessary parameters. Those values were used in a computational example and will hold for the remainder of the paper.

1. The criterion probability of a student not passing minimally acceptable instruction materials or the criterion probability of a student not correctly answering a criterion-referenced test item, p, is .20.
2. The probability of a student not passing instruction or a student incorrectly answering a test item below which false negative errors are critical, p_o, is .10.
3. The probability of a student not passing instruction or a student incorrectly answering a test item above which false positive errors are critical, p_1, is .30. Thus, the indifference region will be $.10 < p < .30$.
4. The risk of committing a false negative error, α, is .01.
5. The risk of committing a false positive error, β, is .10.

Given these parameter values, the decision-making inequality can be computed:

$$s = \frac{-\log \frac{.70}{.90}}{\log \frac{.30}{.10} - \log \frac{.70}{.90}} = \frac{-\log .778}{\log 3 - \log .778} = .186$$

$$I_a = \frac{\log \frac{.10}{.99}}{\log \frac{.30}{.10} - \log \frac{.70}{.90}} = \frac{\log .101}{\log 3 - \log .778} = -1.70$$

$$I_r = \frac{\log \frac{.90}{.01}}{\log \frac{.30}{.10} - \log \frac{.70}{.90}} = \frac{\log 90}{\log 3 - \log .778} = 3.33.$$

Therefore, continue to sample if

$$-1.70 + .186\, n < m < 3.33 + .186\, n,$$

where n is the total number of observations and m is the number of failures or incorrectly answered items.

Figure 1 is the graph of the decision-making inequality. Since $[I_r/(1-s)] = 3.33/.814 - 4.09$, the minimum number of observations necessary

Figure 1
Sequential Decision Making Graph

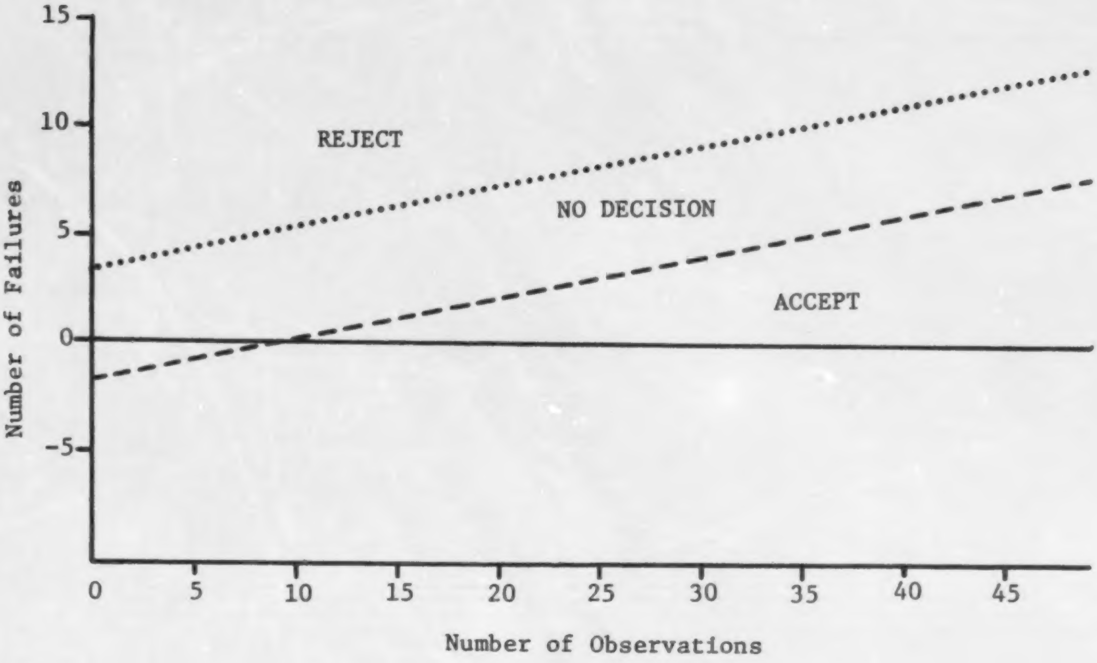
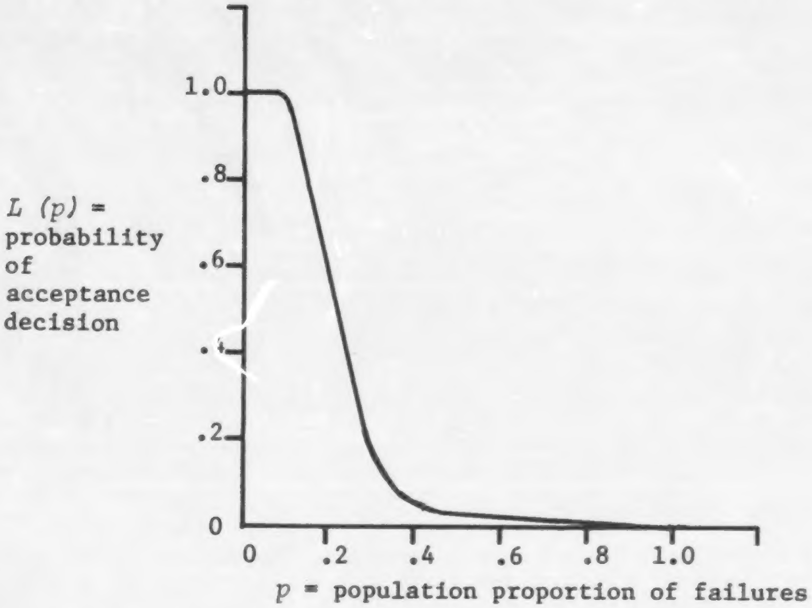


Figure 2
Sequential Testing Operating Characteristic Function



to reject will be five. Furthermore, since $Ia/-s = -1.70/.186 = 9.14$, the minimum number of observations necessary to accept will be 10 .

A graph of the OC function is shown in Figure 2 using the following values:

1. $L(p=o) = 1$
2. $L(p=1) = 0$
3. $L(p=p_o=.10) = 1 - .01 = .99$
4. $L(p=p_1=.30) = .10$
5. $L(p=s=.186) = \dfrac{3.33}{3.33 + 1.70} = .66.$

Figure 3 is a graph of the expected number of observations using the following values:

1. $$E_{p=o}(n) = \frac{\log \frac{.10}{.99}}{\log \frac{.70}{.90}} = \frac{\log .101}{\log .778} = 9.14$$

2. $$E_{p=1}(n) = \frac{\log \frac{.90}{.01}}{\log \frac{.30}{.10}} = 4.09$$

3. $$E_{p=p_o} = .10\ (n) = \frac{.99 \log \frac{.10}{.99} + .01 \log \frac{.90}{.01}}{.10 \log \frac{.30}{.10} + .90 \log \frac{.70}{.90}}$$

$$= \frac{.99 \log .101 + .01 \log 90}{.10 \log 3 + .90 \log .778} = 18.58$$

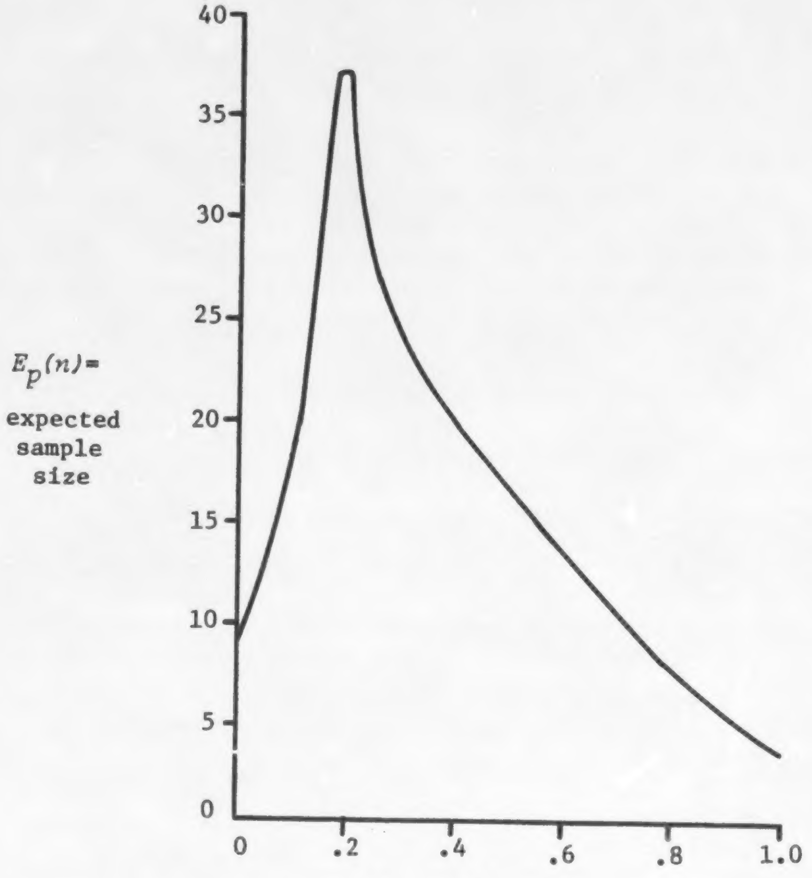
4. $$E_{p=p_1} = .30\ (n) = \frac{.10 \log \frac{.10}{.99} + .90 \log \frac{.90}{.01}}{.30 \log \frac{.30}{.10} + .70 \log \frac{.7}{.9}}$$

$$= \frac{.10 \log .101 + .90 \log 90}{.30 \log 3 + .70 \log .778} = 24.84$$

5. $$E_{p=s} = .186 = \frac{-\log \frac{.10}{.99} \log \frac{.90}{.01}}{\log \frac{.30}{.10} \log \frac{.90}{.70}}$$

$$= \frac{-\log .101 \log 90}{\log 3 \log 1.29} = 37.43\ .$$

Figure 3
Expected Sample Size for the Sequential Test



p = population proportion of failures

Applications

Instructional Materials Data

The first set of example data is a re-analysis of some instructional materials tryout data (Epstein, 1975). It is supportive of the validity of sequential decisions and clearly illustrates the savings possible in the size of the student tryout pool.

Data. The U.S. Army has been heavily involved in the design of audio-visual instruction to teach a wide variety of skills. Tryout data were available for instruction to teach land navigation. The instruction covered eight objectives, each objective having associated with it a performance test that was scored pass/fail. Twenty-eight students participated in the tryout. The data are shown in Table 1.

Table 1
Results of U.S. Army Tryout for Audio Visual
Instruction in Land Navigation, n=28

Objective Number	Number Passing	Percent Passing	95% Confidence Limits for Proportion	
1	27	96	.830	.998
2	26	93	.783	.987
3	26	93	.783	.987
4	25	89	.741	.970
5	26	93	.783	.987
6	26	93	.783	.987
7	20	71	.537	.858
8	16	57	.381	.742

The decision rule used to evaluate this instruction was the 80/80 rule that 80% of the students pass 80% of the objectives. The 95% confidence limits imply that the instruction was certainly acceptable for Objective 1 and that relatively high confidence can be placed in the effectiveness of the instruction for Objectives 2, 3, 4, 5, and 6. The effectiveness of Objective 7 is questionable, and the instruction for Objective 8 is certainly below the minimum requirement.

Results. A sequential testing strategy was applied to the same data. The values for p_o, p_1, α, and β were the same values used in the example discussed earlier. The results of the sequential testing procedure are summarized in Table 2. Since there was no reason to believe that the students were arranged in any particular order, the results from Student 1 in the Army tryout were plotted first, the results from Student 2 second, and so forth.

Table 2
Results of Sequential Testing for
Tryout Data from Army Audio-Visual
Instruction in Land Navigation

Objective Number	Number Tested	Decision
1	10	Accept
2	15	Accept
3	15	Accept
4	20	Accept
5	10	Accept
6	20	Accept
7	17	Reject
8	6	Reject

Average Number Tested = 14.125

The results of the sequential testing procedure agree with the results obtained using the 80/80 rule with the 28 subjects; that is, the instruction was accepted for Objectives 1 through 6 and revised for Objectives 7 and 8. In all cases, fewer students were needed than when using the 80/80 rule. In fact, about half as many students were tested, on the average; and the results for Objective 8 (clearly the objective for which revisions were most needed) were obtained with only six students.

Mastery Decisions

Data. The second set of example data is an application of sequential testing for making individual mastery/nonmastery decisions. It is also a re-analysis of existing data. A total of 237 Military Police students were tested, using the .45 caliber handgun to fire at stationary silhouette targets. Each student fired a total of 240 shots (trials) over a period of two days.

The 240 trials were divided into 3 repetitions of 80 trials each. The first repetition was fired the morning of the first day, the second that afternoon, and the third the following morning. Each group of 80 was subdivided into 10 shots for each of 8 "tables," or distance-position combinations, which define the standard operating procedure for the Military Police Firearms Qualification Course (MPFQC). For this test each group of 10 shots was subdivided into 2 groups of 5 shots. The student had to reload after taking the first 5 shots before shooting the second group of 5 for the same table. The firing order and a description of each table are shown in Table 3.

Table 3
The Military Police Firearms Qualification Course (MPFQC)

Table	Range (meters)	Position	Maximum Time (min.: sec.)
1	35	lying prone	1:45
2	25	standing, no support, preferred hand	1:45
3	25	standing with support, weak hand	1:30
4	25	standing with support, preferred hand	1:30
5	15	standing, no support preferred hand	1:20
6	15	kneeling with support, left hand	1:20
7	15	kneeling with support, right hand	1:20
8	7	crouch	0:24

No feedback was available to a student until after all 8 tables (80 trials) had been fired. Visual sighting by the student of bullet holes in the target was not possible. Holes in the targets were covered with black tape by assistants after each score group of 5-shot trials. During this time, the students had their backs to the targets and were reloading for the next score group.

One 5-round magazine was fired for a given table. While the student reloaded, the score (from 0 to 5 hits) was recorded. Another 5 rounds were then fired, the score was recorded, and the next table was fired. Thus, there were two scores from 0 to 5 for each table.

The data were rescored for sequential testing purposes by dichotomizing each 5-round score according to an 80% criterion (0-3=0; 4-5=1). The new set of 48 dichotomous scores for each student was then considered in the sequence in which they were fired, using the sequential testing procedure with the same parameter values described earlier. Thus, a sequential testing-based pass or fail decision was available for each student. For purposes of comparison, each student's score on the total of the 240 shots was dichotomized (0-191=0; 192-240=1) to yield an overall pass or fail decision.

Results. The 48 scores were insufficient to classify 12 students using the sequential procedure. While it is possible to classify each student on the basis of the probability ratio after all the data have been analyzed, such classifications must be considered as a special case since they alter the previously specified values of the necessary parameters. Table 4 compares the classifications of the 225 students for whom the sequential procedure did lead to a decision with the overall test decision.

Table 4
Sequential and Overall Decisions for the MPFQC

	Sequential Decision		
	Pass	Fail	Total
Overall Pass	45	39	84
Decision Fail	1	140	145
Total	46	179	225

False Positives: 1 or 0.40%
False Negatives: 39 or 17.3%
Total Misclassifications: 40 or 17.8%
Average Number of Items to Reach a Decision: 13.77

The results are encouraging from the point of view of the sample size but both disappointing and confusing with respect to misclassification error. An average of 13.77 observations was required for a decision, which is less than one-third of the total number of 48 observations. Misclassification errors were higher than would be expected from the initial parameters, as well as being in the opposite direction from them. Initially, α, the false negative risk, was set at .01; and β, the false positive risk, was set at .10. The data, however, show a false negative value of .17 and a false positive value of .004. Explanations of these results may lie in violations of the assumptions that (1) the order of item presentation does not introduce bias into the decisions and (2) for any given student, the probability of correctly responding is the same for all items and is equal to that student's hypothetical proportion of correct responses over the entire domain.

Actually, the test could be divided into three subtests. The first eight items in each repetition were similar and relatively difficult, with a mean proportion of hits equal to .64 and a standard deviation of .05. The next six items in each repetition were again similar, but considerably easier, with a mean of .87 and a standard deviation of .04. The last items in each repetition were extremely easy with a mean of .97 and a standard deviation of .002. Thus, each student was faced with a series of difficult items followed by some easy ones, some difficult ones, and so forth. Because of the relatively high passing criterion and the nature of the sequential test, a series of difficult items would tend to favor rejection decisions, particularly if they appeared early in the test--hence, the preponderance of negative decisions observed.

Re-analysis. In an attempt to identify the effect of the inappropriate sequence of items on the sequential test, the data were re-analyzed after partially randomizing the sequence. Reference to a table of random numbers led to the following sequence in terms of MPFQC table numbers: 5, 7, 1, 8, 2, 4, 6, 3. The two scores from each table were again sampled in their original sequence and the above randomized sequence repeated for the three repetitions of the test. The results show some differences.

Although the total number of false negatives decreased, there were increases in the number of students for whom no decision could be made and the average number of observations required. After removing the students for whom no decision was possible, 214 remained. There was again one false positive decision (β observed=.005). Twenty-three false negative decisions (α observed=.107) were observed. The average number of observations increased to 18.39.

The decrease in the number of false negative decisions, along with the increase in the average number of observations and no decisions, tended to favor the conjecture that the original sequence of items led to too many rejection decisions too early. The disappointing results, even after the partial randomization, suggest that the extreme differences in item difficulty severely degrade the applicability of the sequential procedure for data of this type. A clear need for research investigating the robustness of the procedure is indicated.

Performance Testing

Data. The third example examined an application of sequential analysis in the individual Army training and testing context (Knerr & Epstein, 1976). Specifically, the effectiveness of sequential analysis was compared to the total score on the same test as that used for the classification decision and to a criterion external to the test itself. In addition, the efficiency of the sequential testing procedure was measured using the percent of the total number of test items required to make the sequential decisions. The overall objective, then, was to explore the combined effectiveness and efficiency of sequential analysis for the proficiency classification of individuals.

Data for 500 enlisted men at 3 levels of training on 5 combat topics were available from previous research (Knerr, Downey, & Kessler, 1975). In that study, hands-on performance tests were administered approximately one week after the training phase of the research. The topics and number of soldiers tested in each were: Hand Grenades, n=107; Light Antitank Weapon (LAW), n=112; M16A1 Rifle, n=117; Mortar Fire Direction Computer (FDC), n=105; and Surveyed Firing Charts, n=59.

Scores were available for each soldier on each item and on the total test. An index of training was established based on the extent of refresher training that the soldiers received during the research. The lowest level was "no refresher training" (scored 1), the middle level was "some refresher training" (scored 2), and the highest level was "the most effective refresher training" (scored 3). On tests administered immediately after the refresher training, soldiers who received the most effective training scored 82%, on the average, while soldiers who received some training scored 55% correct, on the average.

The sequential analysis procedure was applied to the item data using the parameters described earlier. The items were analyzed in the order that they were administered so that the classification decisions were those that would have resulted if the sequential procedure was applied during the actual test administration. The classification decision for each examinee and the number of items required to reach the decision were noted. In some cases, no decision was made before the total number of test items was exhausted. When no decision was made, the maximum number of test items (total items in the test) was recorded for the examinee. Score codes were 1 for non-proficiency, 2 for no decision, and 3 for proficiency.

Results. The sequential analysis procedure classified 93% of the examinees as either proficient (passed) or non-proficient (failed). No classification decision was made for 7% of the examinees. The Mortar FDC test produced sequential decisions in all cases, and the M16A1 Rifle test produced decisions for all but one examinee. There was no decision reached for between 13-14% of the examinees on the other three performance tests.

Over all 500 examinees, 31% passed and 62% failed according to the sequential decision procedure. It is important to note that the tests on which the sequential decisions were based were administered approximately one week after the training, so that the scores the trained examinees earned were lower than the typical 80% correct scores. In contrast, the scores immediately after training gave an index of the effectiveness of the training and were closer to the typical post-training level. Generally, soldiers who received the most effective training more often passed; and soldiers who received no refresher training more often failed.

The correlations between the sequential decisions and the training index (Table 5) ranged from .20 (LAW test; $p<.05$) to .47 (Hand Grenade test; $p<.01$); the average of the correlations was .34. The correlations between the total test scores and the training index ranged from .27 (M16A1 Rifle test; $p<.01$) to .58 (Hand Grenade test; $p<.01$), with an average correlation of .38. Thus, the sequential decisions were about as effective as the total test scores when effectiveness was assessed against a criterion external to the text.

Table 5
Intercorrelations, Means, and Standard Deviations of
Sequential Decisions, Total Test Scores, and Training Index

Test Topic	Correlation with Training Index		Means and Standard Deviations					
	Sequential Decision	Total Test Score	Sequential Decision		Total Test Score		Training Index	
			Mean	*SD*	Mean	*SD*	Mean	*SD*
Hand Grenades	.47**	.58**	1.26	0.57	15.04	6.10	2.06	0.82
LAW	.20*	.34**	1.55	0.83	14.44	4.86	2.02	0.83
M16A1 Rifle	.25**	.27**	1.69	0.95	41.85	11.75	1.98	0.83
Mortar FDC	.40**	.39**	1.88	1.00	21.48	21.73	2.00	0.82
Surveyed Firing Charts	.35**	.28*	2.46	0.82	25.44	8.55	1.95	0.81

**$p<.01$
*$p<.05$

Since decision accuracy (the extent of classification error) is known to be a problem when the classifications are based on a small number of items, the agreement between the sequential decisions and the total test scores was examined. Where the sequential analysis produced unambiguous proficiency decisions, these decisions agreed with classifications based on the total test score (using the 80% correct criterion) for 88% of the examinees. Thus, the classifications were in error for 12% of the examinees.

Examining just the erroneous decisions, 11% were false positives and 1% were false negatives. These error rates corresponded closely to the predetermined allowable error rates, α and β. The value of α was set at .01, and the extent of false negative decision error was restricted to this amount in these data; β, set at .10, was closely approximated by the obtained error rate for false positives, 11%. Thus, in empirical data on a sample of 500 soldiers, the observed error rates were close to the predetermined ones, indicating that the sequential procedure functioned properly with regard to control over the degree of classification error.

The efficiency of the sequential procedure was examined by comparing the total number of items in each test with the percent of the total number of items required to reach the sequential decisions (Table 6). Averaged over individual examinees, between 19% (Mortar FDC test) and 66% (LAW test) of the total number of test items were required for the sequential decisions. Over the five performance tests, an average of 33% of the items were required to make the sequential decisions. Thus, by using the sequential analysis procedure rather than the traditional procedure of administering the entire test, two-thirds of the items, test administration time, and cost could have been saved. It appears that by combining the accuracy and efficiency results, little accuracy is gained by administering the total test rather than the smaller portion required for the sequential decisions.

Table 6
Items Required for Sequential Decisions
and Test Internal Consistency

	Test Items Required for Sequential Decision		
Test Topic	Average Number	Percent of Total	Internal Consistency (KR20)
Hand Grenades	12.51	43%	.86
LAW	16.41	66%	.83
Mortar FDC	10.43	19%	.99
M16A1 Rifle	15.68	24%	.93
Surveyed Firing Charts	16.02	49%	.96

These data also demonstrated the influence of test homogeneity, or internal consistency, on the sequential decision outcomes. The mathematical basis of the sequential analysis method assumes that the observations are randomly sampled from a single domain. In the present application, the items in a test should all measure the same skill. The order of item administration should be irrelevant, since for any item, passing or failing may indicate the classification decision for the total test. That is, the pass or fail score on any item should correspond to the pass or fail classification decision.

This correspondence implies internal consistency, as measured by the Kuder and Richardson (1937) Formula 20, for which the results are reported in Table 6. The Mortar FDC test had the highest internal consistency (.99), and in the Mortar FDC data the sequential analysis procedure functioned very well. It classified all of the soldiers using only 19% of the test items to reach the decisions, and the sequential decisions were highly related to the training index. Thus, the sequential procedure was accurate and efficient in data that met the homogeneity assumption.

In contrast, the LAW test had lower internal consistency (.83), and in the LAW data the sequential procedure did not work as well. More items were required to reach the decisions (66%), a high number of soldiers had no sequential decision (13%), and the decisions did not relate highly with the training index. In short, when the items did not measure the same underlying skill, or did not measure it reliably, a higher portion of the items was required to reach the decisions; and for a higher portion of examinees, no decision was reached.

The sequential analysis procedure was more effective for tests that measured a single objective and comparatively less effective for tests that either measured diverse objectives or were less reliable. If the scores have low internal consistency for any reason, the order of presentation of the items may have an effect on the decisions. This violates the sequential testing assumption that the observations are a random sample drawn from the same domain.

Conclusions

Wald's sequential probability ratio test for testing the mean of a binomial distribution appears to offer a method for making educational decisions. If the testing situation can be logically interpreted as testing the hypothesis that student performance or instructional effectiveness does not exceed some criterion probability of failure, then the sequential probability ratio test may prove useful. The procedure forces decision makers to explicitly define their definition of acceptable instruction or performance and to consider the misclassification risks that are always present in incomplete sampling. Bringing such issues into the open will certainly be of benefit to the testing community.

The examples included in this paper and other published work (Kriewall, 1969; Linn, Rock, & Cleary, 1972) indicate that the procedure is efficient in terms of the number of observations required for decisions. When the assumptions of the model are met, the procedure appears to be reasonably valid. The major theoretical difficulty lies in insufficient knowledge of the procedure's robustness to violations of the assumptions. On a practical level, the procedure should be easy to implement as part of a computer-managed tested program. It may also prove to be useful (1) in testing situations requiring that students demonstrate their skills at individual stations as part of a performance test; (2) in evaluating instructional effectiveness using small group tryouts; or (3) in other cases where it is practical to consider a decision after each (or perhaps after each small group) of observations. What is now needed, rather than re-analyzing existing data, is experience in using the procedure directly for decision making to determine its proper place in the collection of methods available to decision makers.

References

Baker, E. L., & Alkin, M. C. Annual review paper: Formative evaluation of instructional development. Audio-Visual Communication Review, 1973, 21, 389-418.

Epstein, K. I. Sequential plans and formative evaluation. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC, April 1975.

Knerr, C. S., Downey, R. G., & Kessler, J. J. Training individuals in Army units: Comparative effectiveness of selected TEC lessons and conventional methods (ARI Research Report 1188). Washington, DC: Army Research Institute for the Behavioral Sciences, December 1975.

Knerr, C. S., & Epstein, K. I. Sequential analysis for individual proficiency decisions. Paper presented at the annual meeting of the Military Testing Association, Gulf Shores, AL, October 1976.

Kriewall, T. Applications of information theory and acceptance sampling principles to the management of mathematics instruction (Technical Report 103). Madison, WI: The Wisconsin Research and Development Center for Cognitive Learning, October 1969.

Kuder, G. F., & Richardson, M. W. The theory of the estimation of test rehability. Psychometrika, 1937, 2, 95-101.

Linn, R. L., Rock, D. A., & Cleary, T. A. Sequential testing for dichotomous decisions. Educational and Psychological Measurement, 1972, 32, 85-95.

Novick, M. R., & Lewis, C. Prescribing test length for criterion-referenced measurement (American College Testing Technical Bulletin 18). Iowa City, IA: American College Testing Program, January 1974. [Also in C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), Problems in criterion-referenced measurement (CSE Monograph Series in Evaluation 3). Los Angeles: University of California, Center for the Study of Evaluation, 1974.]

Wald, A. Sequential analysis (2nd ed.). New York: Dover Publications, Inc., 1973. (Originally published, 1947)

ADAPTIVE BRANCHING IN A MULTI-CONTENT ACHIEVEMENT TEST

ROGER J. PENNELL
D. A. HARRIS
AIR FORCE HUMAN RESOURCES LABORATORY
TECHNICAL TRAINING DIVISION

In an environment such as that offered by the Advanced Instructional System (AIS), the potential benefits derivable from adaptive testing become a practical reality. The AIS is an advanced development program to develop a computer-based educational and training system for the Air Force. The heart of the system is a CDC Cyber 70 which currently manages the training process for four courses at Lowry Technical Training Center through the so-called type "A" and "B" terminal. The type "A" terminal is an interactive plasma display terminal with graphic capabilities, while the type "B" terminal has test form reading and scoring capabilities, along with a line printer for issuing student prescriptions. The system is designed to manage the self-paced instructional process of a large number of students who spend approximately 40% of their time in a testing mode. Thus, with a large student flow through AIS courses requiring extensive testing, considerable benefits in terms of reduced training time are potentially available from procedures such as adaptive testing which reduce testing times.

Adaptive testing has also been called branched testing, response-contingent testing, sequential testing, and tailored testing. In the present study, the general term adaptive testing will be used to characterize any attempt to match test items to examinees based on a response history; the goal will be either reducing testing time or obtaining more valid and/or more reliable ability estimates.

Background

Realizing the potential of adaptive testing in a system such as the AIS, the Human Resources Laboratory initiated a multi-phase research study, beginning with the identification of a suitable algorithm to drive an adaptive testing program. The first phase identified the flexilevel approach of Lord (1971a, 1971b) as the tentative algorithm (Hansen, Johnson, Fagan, Tam, & Dick, 1974). Flexilevel testing has a number of advantages over other methods of adaptive testing. Namely, it is easily implemented, it does not require a large item pool, and it theoretically requires only $(n+1)/2$ items (where n is the number of items in the total test pool) to test each examinee. For example, a 25-item test would require only 13 items to test each examinee.

The flexilevel test first administers the item of median difficulty (difficulty levels are ascertained from pretesting). If an item is answered incorrectly, the next easiest unanswered item is given. If an item is answered correctly, the next most difficult unanswered item is given. An examinee continues testing until $(n+1)/2$ items have been answered.

Phase II of the above research effort was conducted in the Inventory Management (IM) course. The Block II test was used for the implementation of this study. The purpose was to validate the flexilevel adaptive testing paradigm with the primary goal of reducing test time. Each student was individually entered into the test, given the flexilevel adaptive test, and then given all remaining items.

A Phase III study was performed in Blocks II and IV of the Precision Measurement Equipment Course (PME). A task analysis was used to group items into five scales and to construct a hierarchy of scales within the test. The intent was to explore the feasibility of adaptive testing both within and across scales.

Study I

Objective

The purpose of the present study was to explore the kinds of conclusions which might be made by simulating flexilevel testing on paper-and-pencil protocols and comparing the results, i.e., estimated parameters, to those data actually collected on the computer terminal (Phase II). The intent was to evaluate the extent to which actual implementation and testing of the model on a computer terminal can be avoided.

A number of simulation studies of adaptive testing have been conducted, e.g., Cleary, Linn, & Rock, 1968a, 1968b; Paterson, 1962; Bryson, 1972; Linn, Rock, & Cleary, 1970. These studies have largely been concerned with ascertaining the potential benefits derivable from an adaptive testing paradigm, rather than extrapolating simulated results to actual adaptive data as this study did. Basically, the question posed by the present study was, "Must one actually conduct a study such as Phase II to ascertain the feasibility of adaptive testing?" Furthermore, "To what extent do simulated results parallel results under actual PLATO testing conditions?"

Method

A sample of 186 paper-and-pencil protocols was obtained from Inventory Management/Material Facilities (IM/MF) Block II. The test was composed of the same items used in the Phase II experiment. The sample was divided into two equal parts: a calibration (C) and a validation (V) sample. The C sample was used to estimate parameters necessary to implement the flexilevel testing algorithm. These parameters were then validated on the V sample in order to evaluate the stability of various dependent measures. The parameters estimated were (1) the item difficulties, implying the item ordering for flexilevel presentation and (2) the regression parameters for converting the flexilevel score into an estimated total score.

The flexilevel score could have been used to make the necessary pass/fail decisions required in a criterion-referenced testing situation such as that found in Air Force technical training. However, for two reasons it was desirable to translate back to the total score metric (percent correct). First, this is the metric traditionally used to assign scores. Second, the extent to which the flexilevel score reproduces the total score is a prime dependent measure in evaluating the feasibility of flexilevel testing.

The flexilevel score was derived as follows:
Let A index the set of items taken under flexilevel testing; and let d_i, $i \varepsilon A$, represent the difficulty of the ith item expressed as percent of the C sample answering correctly;
furthermore, let

$$s_i = \begin{cases} 1 \text{ if item } i \text{ is answered correctly} \\ -1 \text{ if item } i \text{ is answered incorrectly.} \end{cases}$$

Then, the flexilevel score for the jth examinee, on the items completed, was defined as

$$F_j = \sum_i s_i d_i \qquad [1]$$

Stated more simply, F_j was the sum of the item difficulties answered correctly minus the sum of the difficulties answered incorrectly.

Since the total score, X_j, was available as the sum of correct responses divided by the number of items in the item pool (n=25), the usual regression equation

$$\hat{X}_j = a + bF_j \qquad [2]$$

was used to estimate the total score and the associated error $|\hat{X}_j - X_j|$.

It should be noted that the usual flexilevel rule of administering $(n+1)/2$ items to each examinee was departed from in both the Phase II study and the present study. That is, testing for a particular examinee was terminated if he/she were either to take a more difficult item but had already answered all of the difficult items or were to take an easier item but had already taken all of the easy items. This decision rule was used because as a function of entering examinees at varying locations on the item hierarchy, one of the dependent measures was the number of items required to terminate testing.

The dependent variable analyzed in addition to those mentioned above (viz., effect of item hierarchy, variable entry, and error in reproducing total score) was classification error. For a range of criterion levels, the error rate was examined, using $\hat{X}_j$ to classify students as failing or passing relative to their known classification based on X_j.

In addition to the C and V samples, a third sample (N=100) was obtained by randomly selecting test protocols of students who had gone through Phase II testing on the computer. This was possible since at the completion of each flexilevel session (using the same stopping rule described above) all items on the 25-item instrument which had not been administered were given. Thus, complete item protocols were available on this cross-validation (CV) group.

One intention of the Phase II study was to explore the utility of adaptively entering examinees into the item hierarchy. The entry point was calculated using three aptitude tests which the students took before they entered training. It was thought that adaptive entry might further reduce testing time beyond the reduction attributable to taking only $(n+1)/2$ items. Unfortunately, the CV sample was obtained when monitors were having difficulty obtaining the aptitude scores; therefore, the majority of the sample was entered at the $(n+1)/2$th item.

The comparison of the flexilevel results in the CV group, using the parameters estimated in the C group, explored whether or not a feasibility study such as Phase II needed to be conducted. Theoretically, the only difference between the CV and C groups was the use of a computer terminal to administer the test. This assumes item independence in the sense that taking items in a different order would not affect the test score.

Results and Discussion

The item difficulties, along with the correct responses, for the 25 items under study, are presented in Table 1. The mean item difficulty, an estimate of the mean test score, was .804. Typically, criterion-referenced test items tend to be quite easy; however, one of these items was exceptionally difficult (Item 6). Eliminating Item 6 raised the mean to about .84, which indicates that approximately 16% of the sample missed an item of average difficulty. The difficulties in Table 1 implied the ordering of the items for the simulated flexilevel testing; equal item difficulties implied an arbitrary ordering.

Next, the regression parameters for Equation 2 were estimated. Regression estimates for entering the item hierarchy at Item 3, 5, 7, 9, 11, 13, and 15 were calculated. These estimates are presented in Table 2 along with the correlation (validity) between X, the total score, and F, the flexilevel score (see Equation 1). The farther down on the item hierarchy (easier items) students were entered, the more items were required to terminate the flexilevel algorithm. This was vividly displayed by the trend of the regression weights. That is, increasing the entry point reduced the constant term, a, and increased the importance of the b term corresponding to the flexilevel score. The validities beginning at Entry Point 7 were quite good, indicating a high degree of accuracy in predicting total score. However, the cross-validated validities were of more interest.

Table 3 presents the V and CV group validities along with the C group for comparison. It should be noted that $\hat{X}_j$, the estimated total score, was computed using the weights developed in the C group. The validities for the

Table 1
Item Difficulties and Scoring Key, Group C

Item	Difficulty	Key
1	.968	2
2	.936	4
3	.819	2
4	.851	4
5	.809	5
6	.468	5
7	.670	2
8	.819	3
9	.819	1
10	.638	4
11	.915	3
12	.777	4
13	.777	5
14	.862	1
15	.894	1
16	.840	2
17	.840	3
18	.840	5
19	.723	4
20	.862	4
21	.691	4
22	.819	4
23	.755	2
24	.926	1
25	.777	4

Table 2
Regression Weights and Validities, Group C

Entry Point	*a*	*b*	Validity
3	.714	.388	.654
5	.656	.509	.773
7	.617	.560	.847
9	.578	.612	.926
11	.555	.631	.952
13	.524	.661	.972
15	.503	.671	.981

V group were strikingly high. In some cases they were higher than the C group, which indicated that the error in utilizing "non-optimal" regression weights and item difficulties was essentially non-existent. Some shrinkage was encountered in the CV group. However, this shrinkage all but disappeared after Entry Point 11. This indicated that parameters developed on paper-and-pencil protocols cross-validate to results obtained by use of computer terminals for high entry levels.

Table 3
Validities by Entry Point

Entry Point	Group C	Group V	Group CV
3	65	75	60
5	77	78	69
7	85	87	79
9	93	93	83
11	95	95	93
13	97	97	96
15	98	98	98

Note. Decimal points omitted.

Since the items used to construct the flexilevel score were also used (together with additional items) to compute the total score, the validities reported in Table 3 are inflated to some extent. The total score was computed by summing 1's and 0's corresponding to a correct or incorrect item, whereas the flexilevel score was computed by summing weighted item difficulties. Doubtless, the weighted item difficulties have a minimum built-in correlation with the 1-0 protocol.

Table 4
Percent Items Required to Terminate Testing

Entry Point	Group V	Group C	Group CV
3	20	20	19
5	30	30	30
7	41	40	41
9	52	50	52
11	62	60	62
13	70	69	72
15	78	77	80

Table 4 presents the average percent of items needed to terminate the flexilevel algorithm as a function of entry point. For example, when entering at Item 5, all three groups required an average of 30% of the total 25 items (7.5) to terminate the algorithm. The differences between the C sample and the V and CV samples presumably reflect an increase in number of test items required by using non-optimal difficulties, and thus a non-optimal item hierarchy for flexilevel branching. However, this effect was decidedly minimal.

Table 5 presents, in terms of number of items, the average and absolute error made in predicting total score. For example, when each group entered at the 11th item, the estimated total score ($\hat{X}_j$) differed by an average of .9 of an item from the known total score (X_j). Similar to Table 3, these data show comparable results across the three groups entering at Item 11 and above.

Table 5
Item Error in Predicting Total Score

Entry Point	Group V	Group C	Group CV
3	2.0	1.7	1.9
5	1.7	1.5	1.8
7	1.5	1.3	1.4
9	1.2	1.0	1.3
11	.9	.9	.9
13	.7	.6	.7
15	.5	.6	.5

Table 6 shows the average percentage of error of classification across various criterion levels. For a criterion of .70, for example, if $\hat{X}_j \geq .70$ and $X_j \geq .70$ or if $\hat{X}_j < .70$ and $X_j < .70$, the jth student was properly classified. However, if $\hat{X}_j \geq .70$ and $X_j < .70$ or if $\hat{X}_j < .70$ and $X_j \geq .70$, there would have been a classification error relative to the criterion of 70%. The percent of these errors averaged over criterion levels .40, .44, and .96 is the statistic presented in Table 6. When the three groups entered at Item 3, the cross-validated percentage of errors was about 11.5%, which doubtless would be unacceptably high to most course designers. On the other hand, errors of 6 or 7% might be acceptable when balanced against the decrease in overall training time.

Table 6
Percent Misclassified by Entry Point

Entry Point	Group V	Group C	Group CV
3	14	11	12
5	11	10	11
7	10	8	9
9	8	7	9
11	6	6	7
13	5	5	6
15	4	4	5

Conclusions

Making any decision regarding the implementation of adaptive testing involves a tradeoff between potential gains versus potential losses. It has been shown that fairly substantial decreases in the number of test items required are obtainable with very accurate estimation of total score. (An empirical question remaining is whether or not there is a decrease in testing time associated with the decrease in test items.) The tradeoff is relative to the decision categorizing an examinee incorrectly as passing or failing based on a flexilevel score. The above results indicate that this type of error ranges from about 5 to 12%. It should be noted, however, that the criterion used to gauge this error was the total score; this is far from an ideal criterion. What is needed, of course, is the "true score," i.e., the unknown indicator of whether or not a student has accomplished the behavioral objective, which is imperfectly measured by the total test score. Lacking such an indicator, the total score was used; however, there is no reason why the flexilevel test could not be making more valid decisions relative to the "true score." Indeed, this is one of the theoretical benefits attributable to adaptive testing.

The foregoing data have indicated that for reasonably high entry points, parameters estimated from paper-and-pencil test protocols cross-validate remarkably well to groups actually tested at a computer terminal using a flexilevel algorithm. This suggests that feasibility studies running actual subjects may not be called for. Rather, simulated results based on paper-and-pencil protocols may lead to a quick decision regarding whether or not adaptive testing should be implemented.

Study II

Objective

The objectives of Study II were (1) to summarize the data collected under the Phase III contract effort and (2) to offer some conclusions concerning the efficacy of flexilevel testing in an on-going training environment. The analysis was, of course, constrained by the manner in which the contractor implemented the study. However, the present analysis takes a different approach to the data and arrives at slightly different conclusions.

Method

A sample of 133 Precision Measuring Equipment (PME) students who were block tested on the PLATO terminal was obtained. Of those 133 protocols, 61 were Block II tests and 72 were Block IV tests. Both block tests contained 40 items; however, the subject matter covered by the tests was quite different.

A task analysis was performed in order to construct a hierarchical structure for each test. The task analysis grouped items into five relatively homogeneous scales according to item content. The scales were then placed in a hierarchical structure based on the relationships defined by the task analysis.

Table 7

Items Comprising Scales and Difficulties for the Block II Test (Calibration Sample N=105)

Scale 1		Scale 2		Scale 3		Scale 4		Scale 5	
Item	Difficulty	Item	Difficulty	Item	Difficulty	Item	Difficulty	Item	Difficulty
11	.97	24	.97	15	.98	26	.89	34	.95
10	.96	14	.96	29	.94	25	.88	31	.94
6	.96	1	.95	21	.94	39	.88	36	.93
6	.95	5	.90	16	.93	27	.81	37	.90
12	.94	3	.90	20	.92	40	.81	32	.85
7	.92	2	.75	17	.89	28	.70	38	.84
8	.86	23	.74	18	.87			35	.77
		13	.72	19	.85			33	.63
		4	.70	22	.84			30	.51
Mean	.94		.84		.91		.83		.81

Table 8

Items Comprising Scales and Difficulties for the Block IV Test (Calibration Sample N=113)

Scale 1		Scale 2		Scale 3		Scale 4		Scale 5	
Item	Difficulty	Item	Difficulty	Item	Difficulty	Item	Difficulty	Item	Difficulty
15	1.00	1	.96	29	1.00	31	.98	38	.96
16	1.00	10	.90	26	.99	39	.88	4	.95
18	1.00	11	.88	24	.98	37	.88	14	.85
8	.96	5	.88	23	.97	34	.87	13	.84
21	.96	22	.82	25	.94	32	.82	28	.81
2	.92	35	.62	27	.83	33	.70	17	.70
19	.86	7	.61	30	.72	36	.69		
12	.81					40	.57		
20	.82								
3	.67								
6	.58								
9	.58								
Mean	.93		.81		.92		.80		.85

All students entered the test at the median difficulty item of the first scale and were presented items based on the flexilevel algorithm described in Study I. After completing the flexilevel portion of each scale, the students were given the remainder of the items and then started at the median difficulty item in the next scale. This procedure was continued until all five scales were completed.

Results

The items comprising the scales, along with their difficulties, are presented in Table 7 and Table 8. As in Study I, the items were quite easy; the scale mean difficulties ranged from .81 to .94 in Block II and from .81 to .93 in Block IV. The average difficulty of a scale did not necessarily correspond to the position of the scale within the hierarchy. That is, the scales were not ranked in the hierarchy based on average difficulty, but rather by content.

The variables of interest were the proportion correct during the flexilevel portion of the test (S_j) and the flexilevel score (F_j), the latter being modified slightly from Study I. If R is defined as the set of items correct from the flexilevel test, w as the set of incorrect items, and P_i the difficulty of the ith item as obtained from Tables 7 and 8, then

$$F_j = \sum_{i \in R} (1-P_i) - \sum_{k \in w} P_k , \qquad [3]$$

where $i \in R$ and $k \in w$ define the flexilevel score for the jth student. Additional variables of interest were the percent of items saved, the amount of time saved relative to taking the full 40-item test, and the remainder score (the score achieved on those items not taken during the flexilevel portion).

Table 9
Summary Statistics for Dependent Measures

	Block II			Block IV		
Score	Mean	*SD*	*r* with Total Score	Mean	*SD*	*r* with Total Score
Total Score	.85	.39		.82	.39	
S_j (Proportion Correct)	.82	.40	.98	.79	.37	.98
F_j (Flexilevel Score)	.56	.19	.98	.47	.16	.98
% Items Saved	30.4	.89	.96	24.6	.83	.91
Remainder Score	.94	.35	.72	.93	.16	.66

Table 9 contains the means, standard deviations, and correlations with total score for S_j, F_j, percent of items saved, and remainder score for Blocks II and IV. Both S_j and F_j were almost perfectly related to the total score, as evidenced by the correlation of .98. This indicated that after taking about 70% of the items in Block II and 75% of the items in Block IV, the prediction of a student's total score from S_j or F_j was almost perfect.

It was surprising that the relatively crude measure S_j performed as well as F_j, which was intended to be the more sensitive measure. F_j takes into account the difficulty of the item the student takes: Correctly answering an item (i) which is relatively easy results in a relatively small increase in score ($1-P_i$), and relatively large increases occur for correct answers to a relatively difficult item. Incorrectly answering a relatively easy item (i) results in a relatively large decrease in score (P_i), while relatively small decreases occur for incorrect answers to relatively difficult items. However, within the context of the present study, both measures performed equally well.

It can be seen from Table 9 that the mean remainder score was substantially higher than the corresponding total score. This was to be expected; with relatively easy items, students tended to emerge from each scale after taking the most difficult item. Therefore, the remaining items tended to be the easiest items with an associated higher score. Since the items were relatively uniform in difficulty, S_j or F_j should have been a good estimator of the remainder score. In fact, the associated correlations were approximately .55 across blocks.

Two questions remain to be answered. First, can testees accurately be classified into mastery or non-mastery states based on scores (i.e., S_j and F_j) calculated from the smaller item set? Second, was there any actual *time* savings associated with the item savings? The data relevant to the first question are reported in the next section.

Classification analysis. Regression equations for predicting total score ($\hat{X}_j$) from both S_j and F_j were computed (Equation 2). The predicted scores ($\hat{X}_j$) were then compared to the students' observed score (X_j), and the number of correct and incorrect classifications was calculated. For both blocks the course-established criterion of 70% was used to define the cutoff. However, using the total score as the measure of mastery or non-mastery was subject to the same criticism raised in Study I, namely, that the total score is an imperfect measure of mastery, the (latent) trait of interest. The Block II and IV regression equations and classification analyses are presented in Table 10. As can be seen, the prediction of total score pass-fail from either S_j or F_j in Block II was almost perfect; that is, the predicted score ($\hat{X}_j$) misclassified only 1.6% of the sample.

In Block IV F_j classified testees somewhat more accurately than S_j, i.e., 97.2% versus 94.4%. However, the errors in classification based on S_j were

Table 10

Regression Equations and Classification Analysis

Block II				Block IV		
			Regression Equations			
$\hat{X}_j = .08 + .94\ S_j$				$\hat{X}_j = .03 + 1.0\ S_j$		
$\hat{X}_j = .49 + .65\ F_j$				$\hat{X}_j = .48 + .72\ F_j$		

Hit-Miss Analysis Using S_j

Total Score (X_j)	Predicted ($\hat{X}_j$) Pass	Predicted ($\hat{X}_j$) Fail	Total Score (X_j)	Predicted ($\hat{X}_j$) Pass	Predicted ($\hat{X}_j$) Fail
Pass	52	1	Pass	57	4
Fail	0	8	Fail	0	11
% Correct = 98.4			% Correct = 94.4		

Hit-Miss Analysis Using F_j

Total Score (X_j)	Predicted ($\hat{X}_j$) Pass	Predicted ($\hat{X}_j$) Fail	Total Score (X_j)	Predicted ($\hat{X}_j$) Pass	Predicted ($\hat{X}_j$) Fail
Pass	52	1	Pass	60	1
Fail	0	8	Fail	1	10
% Correct = 98.4			% Correct = 97.2		

Table 11

Time (in Minutes) to Complete Scales

	Block II $N = 55$		Block IV $N = 65$	
	Flexilevel	Remainder	Flexilevel	Remainder
1	7.56	3.13	9.14	1.62
2	5.27	0.58	16.25	1.62
3	15.25	2.42	4.05	0.84
4	12.10	1.03	16.48	2.40
5	12.51	1.98	6.83	.93
Total Time on Test	1.03 hrs		1.00 hrs	
Flexilevel Time	.88 hrs		.88 hrs	
Proportion Time Saved	.15 hrs		.12 hrs	

Note. Sample sizes reduced due to occasional computer failure during testing.

conservative, since they classified students as failing the block test when they had actually passed.

Time analysis. The second question, concerning real time savings associated with item savings, was a most critical question. The study by Waters (1975) showed that time savings from adaptive testing procedures are generally minimal; in an operational training environment,a primary concern is whether or not training time and dollars can be saved by adaptive testing.

Data were collected on the amount of time taken by each student to complete the flexilevel portion of the test, as well as the amount of time taken to complete the remainder of the test. These times were collected for each scale in the block tests.

Table 11 presents the mean times for Blocks II and IV. The flexilevel test reduced testing time by only 15% and 12%,respectively. The procedure of starting each student at the median item of each scale required a minimum of 27 items before the flexilevel test was completed. Moreover, as pointed out earlier, those items which were not taken in the flexilevel portion tended to be the easier items and thus were answered relatively faster.

Conclusions

The results of the analyses suggest several conclusions about the efficacy of flexilevel testing in an on-going training environment. First, the proportion correct during the flexilevel test ($S_{.j}$) is as effective in predicting total score as the ostensibly more sensitive flexilevel score ($F_{.j}$). This fact was reflected in the correlation between $S_{.j}$ and total score, as well as in the accuracy of mastery or non-mastery classification. In addition, $S_{.j}$ has the advantage of being in the metric that is most familiar to both students and instructors.

It was also concluded that the modest time savings (12 to 15%) was due to the parameters used to implement flexilevel testing. That is, entering at the median item requires the administration of at least 27 items before exit from the test. In addition, items not taken during the flexilevel test tended to be easier; this was evidenced by the remainder score, which would tend to decrease the time a student needed to complete these items. However, it should be pointed out that even a 15% time saving applied to the large number of students in AIS courses will, in the long run, reflect a significant time savings.

Finally, the selection of the parameters for this study leads to speculation about potentially realizable savings resulting from alternate flexilevel strategies. The following study was designed to investigate that problem.

Study III

Objective

The results of Study II were obviously contingent on the parameters chosen to implement the study. For example, testees always began on the median item of a scale and took all scales. An alternative was to use the flexilevel algorithm at the scale level as well as at the item level (i.e., if a scale were passed, the next hardest scale was taken; if a scale were failed, the next easiest was taken, and so on). Study I has shown that the simulation of the flexilevel algorithm on paper-and-pencil test protocols closely approximated results obtained during testing on a computer terminal. Therefore, using Study II test protocols, the effects of adaptive movement across scales on the various dependent measures was simulated. In addition to implementing the flexilevel algorithm across scales, the simulation considered two other variables. First, the depth or item entry level within a scale was varied in a fashion similar to that used in Study I. Second, this depth notion was extended to the scale level by varying the starting scale between the most difficult and easiest.

Table 12
Items Comprising Scales and Difficulties

Scale 1		Scale 2		Scale 3		Scale 4		Scale 5	
Item	Diff	Item	Diff	Item	Diff	Item	Diff	Item	Diff
Block II									
15	.98	29	.94	5	.90	18	.87	35	.77
11	.97	21	.94	37	.90	8	.86	2	.75
24	.97	12	.94	3	.90	19	.85	23	.74
14	.96	31	.94	17	.89	32	.85	13	.72
9	.96	16	.93	26	.89	38	.84	28	.70
10	.96	36	.93	39	.88	22	.84	4	.70
6	.95	7	.92	25	.88	27	.81	33	.63
34	.95	20	.92			40	.81	30	.51
1	.95								
Mean	.96		.93		.89		.84		.69
Block IV									
15	1.00	1	.96	5	.88	27	.83	36	.69
16	1.00	8	.96	11	.88	20	.82	3	.67
18	1.00	21	.96	37	.88	22	.82	35	.62
29	1.00	38	.96	39	.88	32	.82	7	.61
26	.99	4	.95	34	.87	12	.81	6	.58
24	.98	25	.94	19	.86	28	.81	9	.58
31	.98	2	.92	14	.85	30	.72	40	.57
23	.97	10	.90	13	.84	17	.70		
						33	.70		
Mean	.99		.94		.87		.78		.62

Because of the overlap in item difficulties between the original scales, the items were reordered into scales based entirely on the difficulty indices obtained in the calibration sample. The scales were formed by ranking the items according to difficulty and then forming scales with non-overlapping item difficulties. The position of a scale in the hierarchy was determined by the average difficulty of the scale. Table 12 contains the new scales for the Block II and Block IV tests.

Method

The 133 test protocols obtained during Study II were used as the data in this study. The simulation consisted of varying the levels of three parameters and measuring the effects on the dependent measures. The three parameters manipulated were: (1) scale pass criterion (SPC); (2) scale start (SS); and (3) scale entry level (EL). These are defined below.

Entry level (EL) was used in the same way as in Study I. It defined the item number within each scale where the flexilevel algorithm was started. EL was varied between 1 and 5. If EL=1, the most difficult item was given first; and if EL=5, the fifth most difficult item was given first. EL also defined the minimum number of items that had to be taken before testing within a particular scale was completed. For example, with EL=1 at least one item had to be taken. If it were passed, testing was complete for that scale; if failed, at least one more was taken (the next easiest), and so on.

Scale start (SS) defined the scale within which testing was started, and, thus, took the values 1-5. If SS=5 (the most difficult scale) or SS=1 (the easiest scale), only one scale needed to be taken, i.e., if the most difficult were passed, or the easiest failed, testing was complete.

When the flexilevel strategy was implemented at the item level, the 1-0 item score was used to define the next item to be given, i.e., a "1" implied a more difficult item and a "0" an easier one. In a real sense, this was the criterion for movement between items. In a similar vein, a criterion for movement between scales was needed. This was complicated by variable entry (EL), since EL=1 implied possible scale scores of 1.0, .50, .33, whereas other values of EL implied other ranges of scale scores. Therefore, SPC was not operationalized completely satisfactorily in terms of number of items answered incorrectly. SPC thus was varied between 0 and 3, where a particular value defined the maximum number of items which could be incorrectly answered in order to pass the scale.

The assumption of item independence, which was important in Study I, was also relevant in this study. Namely, a subject taking a particular item in a different order would give the same response as he/she gave in the original order. To the extent that this assumption is true, the results presented below reflect potentially obtainable outcomes from a variety of flexilevel strategies.

Simulations. The computer simulation was used to generate the values of various dependent variables for all possible combinations of the three parameters for both Block II and Block IV. The dependent variables were (1) percent items saved; (2) the percent classified correctly by S_j; (3) the percent classified correctly by F_j; and (4) the correlations with total score for S_j and F_j.

Results and Discussion

Table 13 presents the results of the simulation runs for Block II. Similar to Study I, EL strongly affected the dependent measures. Since EL implied the minimum number of items a student must take, the percent of items saved varied inversely with this parameter, i.e., maximum items saved with minimum EL. Also, as EL increased, the predictiveness of S and F was increased. This was also expected, since as EL increased, the item composite upon which S and F was based increased in size and thus reliability. As predictability increases, the percent of testees correctly classified would be expected to increase. In fact, it did increase.

Table 13
Simulation Results for Block II

Parameter	% Saved	Class (S_j)	Class (F_j)	Correlations $R_{S,T}$	$R_{F,T}$
SPC					
0	67	.933	.932	.829	.840
1	67	.942	.945	.833	.854
2	68	.946	.948	.834	.851
3	69	.919	.942	.830	.845
SS					
1	63	.933	.936	.851	.872
2	61	.949	.948	.877	.893
3	66	.948	.953	.859	.869
4	71	.937	.952	.819	.829
5	80	.908	.921	.753	.774
EL					
1	88	.884	.883	.674	.691
2	77	.925	.934	.817	.833
3	66	.954	.966	.861	.877
4	58	.961	.966	.896	.911
5	50	.949	.961	.911	.925

Note. Results for each parameter are averaged over the values of the other two variables.

The striking aspect of Table 13 is the very large savings in items obtainable with various flexilevel strategies: this is particularly dramatic for EL. At EL=1 only 12% of the items were required to correctly classify nearly 90% of the testees. At EL=2 only 23% of the original items were required to classify over 90% of the testees. This contrasts with the Study II strategy which saved 30% in Block II and 25% in Block IV, while correctly classifying 98% and 96% of the testees, respectively. It was apparent that for only a modest decrease in correct classifications, a large increase in test items saved could be realized. If the relationship between items saved and time saved found in Study I were extrapolated to the present results, a 36% savings in test time could be realized at EL=2.

The relationship of the other parameters to the dependent measures was less clear. SS would be expected to introduce a bow-shaped effect on the dependent variables, since (similar to EL) SS implies the minimum number of scales which must be taken to complete testing. At least three scales are implied by SS=3; SS=2 or 4 implies at least two; and SS=1 or 5 implies at least one. This effect can be seen to some extent in the classification functions and validities which increased to SS=2 or 3 and then decreased. For SPC there was little to choose from in terms of an optimal value. The results for SPC were perhaps idiosyncratic to the generally easy nature of the test items, i.e., varying SPC had minimal implications for all but the least prepared student.

Table 14 presents the simulation results for the Block IV test. Again, EL had the strongest effect on each dependent variable. Indeed, the pattern for Block IV was much the same as the pattern reported for Block II. Results for these blocks suggested that generally optimum values for the parameters were SPC=2, SS=3, and EL=3.

Table 14
Simulation Results for Block IV

Parameter	% Saved	Class (*S*)	Class (*F*)	Correlations $R_{S,T}$	 $R_{F,T}$
SPC					
0	66	.895	.911	.809	.82
1	66	.888	.919	.814	.843
2	69	.886	.915	.818	.847
3	69	.884	.900	.809	.83
SS					
1	63	.887	.908	.823	.858
2	60	.906	.926	.862	.883
3	63	.906	.926	.846	.861
4	69	.894	.917	.812	.829
5	79	.848	.878	.721	.749
EL					
1	88	.853	.862	.639	.656
2	77	.898	.910	.820	.842
3	65	.895	.927	.856	.882
4	56	.895	.925	.868	.897
5	49	.899	.931	.879	.904

Note. Results for each parameter are averaged over the values of the other two parameters.

Table 15 presents the values of the dependent variables for the Block II and IV simulations using the parameter values indicated above. These results indicate that by using approximately 48% of the items there was 100% classification accuracy in Block II and about 93% in Block IV. The correlations of both *S* and *F* with the total score were also quite high. This suggested that total score could be predicted very accurately from either score (a fact observed in the classification data).

Table 15
Simulation Results: SPC=2, SS=3, EL=3

Block	% Saved	Class (*S*)	Class (*F*)	$R_{S,T}$	$R_{F,T}$
Block II	54	1.00	1.00	.94	.95
Block IV	51	.93	.94	.91	.93

Conclusions

Study III has shown that large savings in items and, potentially, test time can be realized through the implementation of alternate flexilevel strategies. The conservative strategy adopted in Study II resulted in only modest item and time savings. However, even these modest savings can result in significant dollar savings when amortized over thousands of technical training students in just one year. Study III has shown that significantly greater savings can be realized with more efficient procedures in the form of optimal values for SPC, SS, and EL.

Conclusions

The overall conclusion from the three studies is that flexilevel testing with variable entry offers an easily implemented testing procedure with potential for significant dollar savings at minimal risk (in the sense of misclassification). Studies I and III, the simulation studies, show the potential power of implementing alternate strategies and the great regularity of the data obtained.

The results from Study I indicate the viability of simulating flexilevel testing on paper-and-pencil protocols to determine optimal entry levels, as well as potential item savings. This type of simulation can be accomplished prior to actual implementation, or the results from Study III can be used directly to guide the selection of an optimal flexilevel strategy.

In any event, it would seem appropriate to implement further flexilevel testing in technical training where the availability of computer terminals permits. For example, since in the Advanced Instructional System students spend 30 to 40% of their time in testing activities, it can be seen that significant training time reductions are potentially obtainable.

References

Bryson, R. Shortening tests: Effects of method used, length, and internal consistency on correlation with total score. Proceedings of the 80th Annual Convention of the American Psychological Association, Washington, DC: The American Psychological Association, 1972, 7-8.

Cleary, T. A., Linn, R. L., & Rock, D. A. An exploratory study of programmed tests. Educational and Psychological Measurement, 1968, 28, 345-360. (a)

Cleary, T. A., Linn, R. L., & Rock, D. A. Reproduction of total test score through the use of sequential programmed tests. Journal of Educational Measurement, 1968, 5, 183-187. (b)

Hansen, D. N. Johnson, B. F., Fagan, R. L., Tam, P., & Dick, W. Computer-based adaptive testing models for the Air Force technical training environment Phase I: Development of a computerized measurement system for Air Force technical training (AFHRL-TR-74-48), Brooks Air Force Base, TX: Air Force Human Resources Laboratory, July 1974.

Linn, R. L., Rock. D. A., & Cleary, T. A. Sequential testing for dichotomous decisions. (College Entrance Examination Board Research and Development Report No. 3, RDR-69-80). Princeton, NJ: Educational Testing Service, 1970. (ETS RB 70-31)

Lord, F. M. The self-scoring flexilevel test. Journal of Educational Measurement, 1971, 8, 147-151. (a)

Lord, F. M. A theoretical study of the measurement effectiveness of flexilevel tests. Educational and Psychological Measurement, 1971, 31, 808-813. (b)

Paterson, J. J. An evaluation of the sequential method of psychological testing. Unpublished doctoral dissertation, Michigan State University, 1962.

Waters, B. K. An empirical investigation of the stratified adaptive computerized testing model. Applied Psychological Measurement, 1977, 1, 141-152.

ADAPTIVE TESTING APPLIED TO HIERARCHICALLY STRUCTURED OBJECTIVES-BASED PROGRAMS

RONALD K. HAMBLETON AND DANIEL R. EIGNOR
UNIVERSITY OF MASSACHUSETTS, AMHERST

Objectives-based instructional programs (e.g., Glaser & Nitko, 1971) were introduced to provide instructional programs that would be maximally adaptive to the needs of individual learners. While the specific methods of implementation have varied widely, common to all has been the notion that a curriculum should be defined by a set of objectives. Another common theme of objectives-based programs has been that student progress should be measured by comparing student performance to standards of performance set on the objectives defining a curriculum; student progress was not to be measured by comparing the performance among students (Hambleton, 1974).

Criterion-referenced tests were introduced initially by Glaser (1963) and Popham and Husek (1969) to provide a way for collecting the kind of information needed to assess student performance relative to a set of objectives. More recently there have been numerous contributions to the emerging field of criterion-referenced testing technology (e.g., Glaser & Nitko, 1971; Hambleton & Novick, 1973; Hambleton, Swaminathan, & Algina, 1976; Millman, 1974; Popham, 1975).

Student mastery of objectives in a segment of a curriculum is often determined by an administration of a criterion-referenced test. "Mastery" is inferred when a student's test performance on a set of items measuring an objective exceeds some minimum performance level. The minimum performance level for mastery is often referred to as a cutting score or passing score.

In theory, criterion-referenced test scores can be made as reliable and valid as necessary by adding additional test items. Unfortunately, making a mastery--non-mastery decision on each of the objectives measured by a criterion-referenced test often requires a considerable amount of testing time. Therefore, it is usually impractical to consider lengthening tests, particularly to the length that would often be necessary to accomplish some desired goal for reliability and validity of test scores.

Some critics have argued that there is already too much criterion-referenced testing in objectives-based programs. On the other hand, some increase in testing time can be defended on the grounds that test response data is closely tied to the objectives defining the curriculum and that the data are used to monitor student progress. Nevertheless, it seems clear that research is needed on procedures offering potential for reducing testing time without reducing the quality of decision-making from test score results.

The use of Bayesian statistical procedures represents one promising method for reducing testing time and/or improving the quality of mastery decisions (Hambleton & Novick, 1973; Novick & Jackson, 1974; Swaminathan, Hambleton, & Algina, 1975). This method is particularly appealing because it requires no change from the most common methods of test administration. Improvements in decision making are attributable to the utilization of information ignored by non-Bayesian procedures. Bayesian procedures may use not only the *direct information* provided by an examinee's test score, but they also make use of *collateral information* contained in the data of other examinees and of *prior information* on other relevant data that are available on the examinee (e.g., test scores from other segments of the course).

In one simulation study Hambleton, Hutten, and Swaminathan (1976) compared several Bayesian estimation procedures with several classical procedures for assessing student mastery and making instructional decisions. They reported modest gains from use of the Bayesian estimation procedures. On the negative side, Bayesian statistical procedures are based on restrictive assumptions, and robustness of the procedures has not been studied extensively. Also, some individuals feel that the utilization of group information to influence individual mastery estimates is a contradiction of one of the fundamental postulates of objectives-based instruction, that is, each student is judged on his/her own merits; thus, mastery decisions should not depend on the performance of other students.

A second promising solution to the testing time problem is offered by adaptive testing. Adaptive testing has been defined as a strategy for testing in which the sequence and number of test items a student receives are dependent upon his/her performance on earlier test items. Of special interest in this paper is the application of adaptive testing to hierarchically structured objectives-based programs. When the hierarchy of objectives is specified, inferences can be made about student mastery of objectives in the hierarchy which have not been tested. If, for example, a student is tested and found to have "mastered" a particular objective, all prerequisite objectives can also be considered to have been mastered. If an examinee has "not-mastered" an objective, it can be inferred that all objectives to which it is a prerequisite are also unmastered.

A considerable amount of work on adaptive testing has been done since the late 1960s (e.g., Lord, 1970; Weiss, 1977; Wood, 1973). Much of the research has concentrated on adaptive testing schemes for improving the precision of measurement of examinee ability while decreasing the amount of testing time. A second problem area (not independent of the first) is one of classifying examinees into "mastery" and "non-mastery" states for a set of objectives that can be arranged hierarchically.

An additional problem of considerable importance is to find optimum adaptive testing strategies for assigning examinees to "mastery states" for the objectives in a learning hierarchy. (The expression "learning hierarchy" will be used in this report to refer to a set of objectives arranged into a hierarchy reflecting dependencies among the objectives.) A further problem concerns the amount of testing time that can be saved in comparison with testing examinees on all of the objectives included in a learning hierarchy.

There are two areas requiring attention from researchers before results of empirical studies on adaptive testing will be of much value. The first area is the construction and validation of criterion-referenced tests; the usefulness of adaptive testing is related directly to the validity of criterion-referenced tests. Certainly, unless the intended interpretations of the criterion-referenced test scores are validated, decisions or descriptions based on the test data are suspect. The second area is the construction and validation of learning hierarchies. If information about a hierarchy of objectives is to be used to influence the adaptive testing plan, it is essential that a learning hierarchy be produced and validated.

Three topics will be covered in this paper: (1) construction and validation of criterion-referenced tests, (2) construction and validation of learning hierarchies, and (3) adaptive testing research.

Construction and Validation of Criterion-Referenced Tests

Since 1969 over 400 papers have been written on the topic of criterion-referenced testing. Unfortunately, this avalanche of papers reflects almost as many views and assumptions as there are contributors to the literature; and as a result, there has been substantial confusion in the field. However, with the important integrating work of Glaser and Nitko (1971); Millman (1974); Harris, Alkin, and Popham (1974); and Hambleton, Swaminathan, Algina, and Coulson (1978) much of the terminology has been standardized, many issues delineated, and many important technical matters resolved. It is now known how criterion-referenced tests are developed, test scores used, and test score reliability and validity assessed. Although many problems still remain, at least it is now possible to initiate a successful criterion-referenced testing program by drawing on a more than adequate testing technology.

A criterion-referenced test can be defined as a test constructed to permit the determination of an examinee's level of mastery relative to a "well-defined" behavior domain (Popham, 1975). Each behavior domain is keyed to an objective in the learning hierarchy. A criterion-referenced test is constructed to assess student mastery of each objective in a learning hierarchy. On many occasions, items from different behavior domains are included in the same test. In such cases, examinees receive scores based on their responses to items in each behavior domain. The definition of a criterion-referenced test is not unlike the definition offered by Millman (1974) for *domain-referenced tests*.

Preparation of Objectives

To operationalize the skills included in a learning hierarchy, the usual strategy is to write "behavioral objectives." Behavioral objectives have some desirable features, e.g., they are easy to write. Unfortunately, a behavioral objective usually lacks the clarity necessary to permit a clear determination of the domain of test items which measure the behaviors defined by the objective. If the proper domain of test items measuring an objective is not clear, it is impossible to select a representative sample of test items from that domain.

Current thinking among criterion-referenced test theorists is that objective statements should be expanded into domain specifications (Popham, 1975). Domain specifications are intended to reduce confusion among test users and item writers over the set of behaviors spanning the area defined by an objective. Improved clarification of domains of items measuring objectives will contribute substantially to the improvement of criterion-referenced test construction methods.

Popham outlined four steps for the development of domain specifications. First is the preparation of a general description; it could be a behavioral objective, a detailed description of the objective, or a short cryptic descriptor. Second, a sample test item is prepared. This step will help clarify the proper domain of test items and specify item format. In the third step, perhaps the most difficult, it is necessary to indicate the behaviors included in the domain. At this stage, the writer of a domain specification could list exemplary behaviors as well as behaviors which are not included in the domain specification if there might be some confusion over their status. In the fourth and final step, characteristics of response alternatives are specified.

The important aspect of these four steps is that they are *specific*. It is not necessary, however, that they specify a homogeneous set of behaviors. Domain specifications for objectives in a learning hierarchy will typically be more homogeneous than those written for other purposes. If domain specifications for objectives become too heterogeneous, it will become considerably more difficult to sequence objectives in a hierarchy.

Excellent examples of domain specifications are those prepared by Hively, Patterson, and Page (1968; see also Millman, 1974; Popham, 1975). Their work is based on two requirements: (1) All items which could be written from the domain to be tested must be written (or known) in advance of the final item selection process, and (2) a random or stratified random sampling procedure must be used in the item selection process.

One way to achieve these two requirements is through item forms analysis. An item form has the following characteristics:

1. It generates items with a fixed syntactical structure.
2. It contains one or more variable elements.
3. It defines a class of item sentences by specifying the replacement sets for the variable elements.

One of the obvious advantages of such a system is that the workload which would be required in writing the larger number of items needed to satisfy the two conditions above would be reduced.

The Hively et al. study is important in that it demonstrated that it was possible to develop and use item generation rules to construct a test. The study also underscored one of the major weaknesses of item generation procedures--the procedures are more easily employed with highly structured subject matter areas, such as mathematics. Although it is unlikely that the clarity of domain specifications offered by Hively and his colleagues can be produced in very many content areas, clarity is still necessary if criterion-referenced tests are to be useful for assessing mastery within an adaptive testing scheme.

The importance of having domain specifications cannot be overemphasized. What is typically desirable is to be able to interpret an examinee's test score as an estimate of an examinee's performance level in the larger domain of items measuring an objective. In addition, the scores are used to make instructional decisions (assign examinees to "mastery" and "non-mastery" states on each objective in the learning hierarchy). When the domain of items is unclear (or unspecified), only a "weak" criterion-referenced interpretation is possible, i.e., examinee test performance must be interpreted in terms of the items included in a test. Generalizations about student performance to a domain of behaviors, i.e., a "strong" interpretation, are not possible.

Item Writing

This step is not very different from writing test items for norm-referenced tests. There are a set of principles for item writing that should be followed; it is necessary, however, for item writers to attend to the domain specifications. Test items should "tap" behaviors in the domain of behaviors defined by domain specifications.

Item and Test Score Validity

Cronbach (1971), Messick (1975), and Linn (1977) have argued that to validate interpretations of criterion-referenced test scores (i.e., to determine what is being measured), it is necessary to proceed beyond a consideration of content validity. Until recently, it was thought that content validity considerations of a criterion-referenced test were sufficient. However, Messick (1975) stated:

> The major problem ... is that content validity ... is focused upon test *forms* rather than test *scores*, upon *instruments* rather than *measurements*. Inferences in educational and psychological measurement are made from scores, and scores are a function of subject responses. Any concept of validity of measurement must include reference to empirical consistency. Content coverage is an important consideration in test construction and interpretation, to be sure, but in itself it does not provide validity. (p. 960)

Content validity is a test characteristic. It does not change from one group of examinees to another. However, the validity of test score interpretations *will* vary from one testing situation to another; therefore, construct validation studies must be conducted. For example, if a criterion-referenced test is administered by mistake under highly speeded testing conditions, the meaning of the test scores obtained from the test administration will be different than if the test had been administered with more suitable time limits.

In preparing criterion-referenced tests for use in adaptive testing, both the content validity of the test and the construct validity of the test scores must be established. (For a lengthy discussion of the validity question, see Hambleton, 1977.)

Content Validity

Content validity of a criterion-referenced test is studied by investigating two questions:

1. Is the domain specification clear?
2. Is there agreement that a set of items adequately samples the behaviors defined in the domain specification?

The first question can be studied by comparing test items generated by different item writers and analyzing the judgments of content specialists about items relative to the domain they were developed to measure. Three techniques for the collection and analysis of the judgments of content specialists have been described by Rovinelli and Hambleton (1977).

The second question is difficult to investigate, unless the domain of items is completely revealed in a domain specification. The question can be investigated by Cronbach's (1971) interesting, but somewhat impractical (at least for small-scale test development projects), duplication experiment. The judgments of content specialists are also useful.

The matter of technical quality of test items is handled at the test development stage. It could be done at the content validation stage also by asking content specialists to address the matter of technical quality of items in their review.

Construct Validity

According to Messick (1975), the definition of construct validation is "the process of marshalling evidence in the form of theoretically relevant empirical relations to support the inference that an observed response consistency has a particular meaning"(p. 955).

Construct validation studies begin with a statement of the intended use of the test scores; a statement of use will provide direction for the kind of evidence that is worth collecting in a construct validation study. Following is a description of some kinds of investigations that can be conducted to study construct validity of a criterion-referenced test.

Item statistics. When items in a domain are expected to be relatively homogeneous, a test developer can compare estimates of item difficulty parameters, item discrimination parameters, or both. Since items from a homogeneous domain of items measuring an objective would be expected to have similar item parameters, estimates of the parameters can be compared in order to detect items that deviate from the norm defined by the remaining items. Such "deviant" items can be carefully scrutinized for flaws that may be reducing their validity.

Instructional experiments. One solution, albeit costly, is designing experiments to determine the construct validity of criterion-referenced test scores. Individuals are randomly assigned to one of two groups. One group receives instruction on content defined by a domain specification. The other group receives no instruction. If the treatment is effective, and this could be determined from past experience, higher test scores by the experimental group would support the construct validity hypothesis. Experiments on other objectives could also be conducted, but it would be desirable to alternate control and experimental groups so that no single group of examinees would be denied instruction.

Factor analysis. Factor analysis is a commonly employed procedure for the dimensional analysis of items in a norm-referenced test or scores derived from different norm-referenced tests. It could be used in construct validation studies of criterion-referenced test scores. Perhaps one reason for its lack of use is that the usual input for factor analytic studies is correlations. Correlations are often low between items in a criterion-referenced test or between criterion-referenced test scores and other variables because criterion-referenced test-item and test-score variability is often not very great.

The problem can be remedied by choosing a sample of examinees that includes a range of scores. For example, a test developer should choose a sample of examinees that includes both masters and non-masters of the material measured by the test of interest. The research problem, in the language of factor analysis, becomes a problem of determining whether or not the factor pattern matrix has a prescribed form. If the intercorrelations among a set of items in a criterion-referenced test are factor analyzed, as many factors would be expected to be obtained in the factor solution as there are objectives measured by the test. Test items should "load" only on the factor (or objective) that they were designed to measure. Items deviating from this pattern would be carefully studied for flaws.

Experimental studies of sources of invalidity. There are numerous sources of error that can reduce the validity of a set of criterion-referenced test scores,and the influence of many factors is involved. Some relevant questions are:

1. How clear were the test directions?
2. Was there confusion in using the answer sheets?
3. Was the test administered under speeded testing conditions?
4. Were the examinees motivated to do their best?

To the extent that any of these (and many other) factors influence test scores, the descriptive interpretation of the test scores is weakened.

Required are experimental studies of potential sources of error to determine their effect on test scores. Results of these studies can be used to further clarify domain specifications. For example, if it is discovered that item format influenced test scores, the item type to be used could be specified after it is determined which item produced the most construct valid test scores.

Item Selection and Test Length

Items should be selected to be representative of the domain of items measuring the objective of interest. A random or stratified random sample of items is essential to permit generalizations from examinee test scores on the sample of test items to the larger domain of items.

How many items are needed in a test to measure each objective? When using criterion-referenced tests to assign examinees to mastery states (i.e., "mastery" and "non-mastery"), the problem of determining test length is related to the size of misclassification errors one is willing to tolerate. One way to assure low probabilities of misclassificationis to make tests very long, but this is impractical.

Millman's binomial test model. Millman (1973) considered the error properties of mastery decisions made by comparing an observed proportion-correct score with a mastery cut-off score. By introducing the binomial test model, it is possible to determine the probability of misclassification conditional upon an examinee's true score, an advancement score, and the number of items in the test. (The advancement score is the minimum number of items that an examinee needs to answer correctly to be assigned to a mastery state, as distinguished from the cut-off score,which is the point on the true mastery or domain score scale used to sort examinees into mastery and non-mastery states.)

By varying test length and the advancement score, it is possible to determine the test length and advancement score that produce a desired probability of misclassification for a *given* true score. In Millman's model, the assumption is made that examinees attempt each item in the test measuring an objective with probabilities equal to their "true scores" for the objective. ("True score," "domain score," or "level of functioning score" is the proportion of items in the domain of items measuring an objective that the examinee can answer correctly.)

It follows, then, that the probability of an examinee with true score p obtaining a test score x on an n-item set of items measuring an objective is given by

$$\text{Prob } (x|p) = \binom{n}{x} p^{x}(1-p)^{n-x} \quad . \qquad [1]$$

Let n=5 and p=.60. The probabilities of an examinee obtaining a score of 4 or 5 are .26 and .08, respectively. To calculate the probability of a "false-positive" error when both the advancement score and cutting score are set at 80%, it is necessary to calculate the probability that the examinee's test score equals or exceeds 80%. For an examinee with true score equal to .6, the probability of a false positive error is .34. In the same way, the probability of misclassifying an examinee (false positive error or false negative error) can be calculated for any examinee with known true score on a test of n items with known advancement score and cutting score.

Of course, in practice, examinee true score is unknown; in fact, it represents the characteristic of an examinee being estimated. Nevertheless, experience with Equation 1 can be very helpful in determining acceptable test lengths. Useful tables based on the binomial test model for various test lengths, true scores, advancement scores, and cutting scores have been reported by Millman (1973).

Adaptive sequential probability models. Millman's work is particularly helpful to classroom teachers and other criterion-referenced test developers. If, however, a computer terminal for test administration is available, more suitable solutions to the test length problem can be obtained. Ferguson (1969) used Wald's (1947) sequential probability ratio test to assign examinees to mastery states. Ferguson's procedure allows the tester to vary the test length for each examinee. Test length is varied to insure that a mastery—non-mastery decision for each examinee can be made so that the risks of a false positive or false negative error are below values set by the tester.

The tester specifies a minimum cutting score, p_0, that an examinee must meet to be assigned to a mastery state. Also, a second cutting score, p_1, where $p_1 < p_0$, is specified so that an examinee whose score is less than p_1 is assigned to a non-mastery state. Next, the tester specifies the probability of making a Type I (α) error (a false-negative error) and the probability of making a Type II (β) error (a false-positive error) that he/she is willing to tolerate.

If $\hat{p}$ is the proportion of items that an examinee has answered correctly, there are three possible mutually exclusive decisions:

1. If $\hat{p} \geq p_0$, classify the examinee as a master.
2. If $\hat{p} \leq p_1$, classify the examinee as a non-master.
3. If $p_1 < \hat{p} < p_0$, do not make a decision; administer more test items.

More formally, the sequential probability ratio test of strength (α, β) for testing

$$H_0 : p \geq p_0$$
$$H_1 : p \leq p_1$$

can be viewed as a Bernoulli-type experiment, which means results fit the binomial distribution. Here, p is the true score for the examinee. The risks of misclassification are as follows: The probability of a false-negative error ($\hat{p} \leq p_1$, $p \geq p_0$) should not be greater than α and the probability of a false-positive error ($\hat{p} \geq p_0$, $p \leq p_1$) should not be greater than β.

Once p_0, p_1, α, and β are specified for any particular objective, it is possible to prepare a table indicating the number of items that must be passed (or failed) at test lengths from 1 to n items in order to lead to mastery decisions (or non-mastery decisions). Formulas for preparing a table are offered by Wald (1947); applications of the formulas to adaptive testing are provided by Ferguson (1969).

Reliability

What form should reliability take for a criterion-referenced test? Suppose instructional decisions are made by comparing examinee test performance to a minimum proficiency level or cutting score for each objective. Of interest is the consistency of "mastery" and "non-mastery" decisions across parallel-forms (or repeated administration) of the criterion-referenced test. An intuitively appealing measure of agreement between the decisions made (for each objective) on the two administrations is $p_{11} + p_{22}$, where p_{ii} is the proportion of examinees placed in the i^{th} mastery state on each test administration. However, this measure of agreement does not take into account the agreement that could be expected by chance alone and hence does not seem entirely appropriate.

The coefficient kappa, designated κ (Swaminathan, Hambleton, & Algina, 1974), takes into account chance agreement and thus appears to be somewhat more appropriate for use with criterion-referenced tests. Coefficient kappa (the expression for reliability or consistency of mastery decisions on an objective) is defined as

$$\kappa = (p_o - p_c) / (1 - p_c), \qquad [2]$$

where p_o, the observed proportion of agreement, is given by

$$p_o = \sum_{i=1}^{2} p_{ii}; \qquad [3]$$

and p_c, the expected proportion of agreement, is given by

$$p_c = \sum_{i=1}^{2} p_{i.}\, p_{.i}\ . \qquad [4]$$

Both $p_{i.}$ and $p_{.i}$ represent the proportions of examinees assigned to the mastery state i on the first and second test administration, respectively.

Since p_o is the observed proportion of agreement and p_c is the expected proportion of agreement, kappa can be thought of as the proportion of agreement that exists over and above that which can be expected by chance alone. Kappa provides a useful indicator of decision-making consistency. If it is lower than desirable, test length should be increased. When Wald's sequential probability ratio test is used, reliability or consistency of decision making across a group of examinees for particular objectives can be replaced by the values of α and β chosen for each objective.

Construction and Validation of Learning Hierarchies

One of the most promising lines of research to come out of learning theory investigations over the past fifteen years is the concept of learning hierarchies. Starting with a study by Gagné (1962), the research in this area has progressed until at present there is a well-defined methodology for the development of learning hierarchies and statistical tests for validating posited hierarchies. The state of the field is developed sufficiently so that it can be easily linked to related fields, such as adaptive testing.

Gagné (1970) used the term "learning hierarchy" to designate the set of dependencies among component skills or objectives within a learning task. He also suggested that learning hierarchies might define optimal sequences for presenting learning events. Clearly, besides the instructional process, establishing learning hierarchies has some very useful implications for the testing process. The objectives identified in a learning hierarchy can be measured using the test development and validation methods outlined in the previous section. An adaptive testing scheme could be used to determine which criterion-referenced tests would need to be administered in order to determine an examinee's mastery states on objectives in a learning hierarchy.

Generating Learning Hierarchies

Most researchers investigating learning hierarchies develop a provisional ordering of the instructional objectives comprising the hierarchy before initiating any type of validation procedure. At present there are two possible ways of developing provisional orderings of objectives in a hierarchy. Passmore (1974) aptly named these two methods "introspection" and "statistical fishing."

Gagné (1962) developed the logical questioning technique called "introspection." The researcher takes the final objective and asks, "What would an individual need to know to display competence in this subject matter?" The same question is then applied to each of the behaviors specified by the last application of the questioning techniques. This procedure is continually used until simple behaviors are reached which cannot be linked further to other necessary pre-behaviors. The generated sequence can then be represented as a hierarchy, which at this point can be considered only as provisional. There will be general behaviors at the top of the hierarchy and more specific, subordinate behaviors near the bottom.

Almost all of the learning hierarchy research done to date has proceeded from a provisional hierarchy developed in this particular fashion. The method presupposes that the individual doing the questioning is well acquainted with the subject matter and is capable of relating necessary pre-behaviors to the considered terminal behavior. This is critical because, as will be discussed later, the methodology for validating hierarchies can only lead to acceptance or rejection of a posited hierarchical connection. It cannot generate hierarchical links that have not been provisionally specified. Thus, it is critical that an individual or individuals well acquainted with the domain be involved in the introspection process.

"Statistical fishing methods" for developing hierarchical clusters, which could be applied to learning hierarchy research, have been advanced but to date have seldom been utilized. This would appear to be for two reasons. First, the methodology requires the understanding of statistical procedures not usually directly in the researcher's command. Second, the available literature has not been directly related to the establishment of learning hierarchies, thus requiring the researcher to link two fields not presently related. For instance, much of the work done has used measures of similarity between examinees, while the work on hierarchies would require some type of measure involving objectives.

The available procedures of a statistical nature involve either the application of statistical clustering techniques, such as discriminant analysis and hierarchical cluster analysis (Tryon & Bailey, 1970) or application of hierarchical clustering schemes based on principles of numerical taxonomy (McQuitty, 1970; Johnson, 1967). The former methods either yield final clusters with no hierarchical structure or depend upon an a priori definition of the number of clusters. It would seem, therefore, that the methods related to numerical taxonomy would be preferred.

The iterative procedures available relate similarity measures on individuals at each iterative stage until, at the final stage, all individuals are related and belong to one overall cluster. Researchers must make a decision as to which iterative stage they want to view the hierarchical structure. The Baker (1972) article contains an excellent example of the use of Johnson's (1967) MAX procedure, in which the operation of the clustering algorithm can be seen. This example and others are based upon individual similarity measures, but the procedures appear to be adaptable to usage involving similarity measures on objectives.

Evaluating Provisional Hierarchies

Once a provisional ordering of objectives in a learning hierarchy has been advanced, the ordering is evaluated to determine whether the connections between objectives should be rejected or accepted. It is at this level of hierarchy development that extensive research has been done. An extensive review of methods for validating learning hierarchies was prepared by White (1973), who then (White, 1974a, 1974b) developed a new model for validating learning hierarchies that appears to be the best possible procedure to use at present. Following is a brief explanation of the steps involved, taken and adapted from White (1974a):

1. Define the top objective of the hierarchy.
2. Derive, using Gagné's method of questioning, the subordinate objectives, being careful not to include verbalized knowledge objectives.
3. Check the postulated hierarchy with subject matter specialists.
4. Subdivide the objectives, if necessary, so that clear definitions are obtained.
5. Using a sample of students, check that objectives in the hierarchy are distinct (see White, 1974a).
6. Prepare instructional materials for the objectives with the test items (two or more for each objective) to be administered following instruction on the objectives.
7. Have a suitable number of students (at least 150) work through the instructional materials.
8. Analyze the results to see whether the postulated connections can be rejected, using the statistical test developed by White and Clark (1973).
9. Remove from the hierarchy all connections for which the probability of a hierarchical connection is small.

White field-tested his procedure (White, 1974a) as did Linke (1975); both found the procedure to work well with the objectives being taught and tested. It is noteworthy that these investigations and those of Gagné and others (see White, 1973) all involve hierarchies in the areas of science and mathematics. Possible reasons for this fact may be twofold: One, the method of questioning about necessary prerequisite skills may not work well for the social science and language fields; perhaps some sort of statistical clustering method needs to be employed. Two, until recently, the problems of writing higher order objectives for the social science and language fields may have proven to be a hindrance. Whether the methodology outlined by White can be applied to areas other than math or science remains at present a research question of considerable importance.

Statistical Techniques for Validating Hierarchies

As discussed in the previous section, a statistical test is desirable for ascertaining whether two objectives are connected in a linear fashion. Researchers who did the early work in the field of learning hierarchies used a number of different indices, all of which suffered to a greater or lesser degree from the following two maladies: (1) The indices can have values that indicate a hierarchical connection, even when the objectives are independent, and (2) the procedures are deterministic and do not allow for statistical tests of fit. Noteworthy of the indices falling into this category are Gagné and Paradise's (1961) "proportion positive transfer," Resnick and Wang's (1969) use of Guttman's coefficient of reproducibility, and Capie and Jones' (1971) use of the phi coefficient.

White and Clark (1973) developed a statistical test that can be used when more than one test item is used to measure each objective. A hierarchical relationship is postulated. Then the number of examinees is observed who answer all the test items for the lower objective incorrectly and answer all the test items for the higher objective correctly. The connection is judged invalid when this number of examinees exceeds a critical value, specified for the probability of wrongly rejecting the null hypothesis that the connection is hierarchical. Examples of the procedure are discussed in White and Clark (1973) and White (1974a).

While the statistical distribution theory is somewhat complicated, the practical application (based in part on the use of power functions) is straightforward. A cogent word of caution has been advanced by Passmore (1974), having to do with the relationship between sample size and the power of significance tests. Passmore has advised that valid connections may be rejected if standard hypothesis testing procedures are used with the large samples White has suggested. Passmore has further indicated that the relationship of White and Clark's power function to sample size considerations should be more fully explained to practitioners.

Recently, Dayton and Macready (1976) have developed a more general set of statistical procedures for validating hierarchies that subsume White and Clark's work as a component. This procedure can be used to test a whole hierarchy, whereas all previous methods test connections between pairs of objectives in a hierarchy. The model offers a χ^2 goodness of fit test of observed proportions data to expected proportions generated on the basis of the hypothesized hierarchical relationship. Furthermore, the model allows for the testing of arbitrary hierarchies, which include linear and branching patterns; and it allows for concept attainment models, in which the hypothesized pattern vector consists of ones or zeros to specify mastery or non-mastery of objectives in the hierarchy.

Dayton and Macready have offered two reasons why they feel that their probabilistic model is preferable to use when testing the connection between two objectives hypothesized to be hierarchical, rather than White and Clark's test. First, the estimation procedure of White and Clark is not maximum likelihood, while Dayton and Macready's method is; the Dayton and Macready method is thus more desirable from the point of view of sampling properties. Second, using the Dayton and Macready procedure, a more general test of inclusion can be performed. Two sets of pattern vectors can be developed--one implying inclusion or a hierarchical connection and the other having the pattern vector for inclusion augmented by the discrepant pattern configuration, which demonstrates a non-hierarchical relationship. Then a χ^2 test can be performed on the difference between the expected proportions generated from the two solutions of the model. A significant χ^2 on the difference would indicate a non-hierarchical connection.

The model and statistical procedures developed by Dayton and Macready represent a significant development in the validation of learning hierarchies. In addition to the facts that the model and statistical test can be used to validate an entire hierarchy and that there are available computer programs to use, the procedure has one important advantage not offered by White and Clark's procedure. While an a priori hierarchy may yield an insignificant χ^2 when the expected proportions generated are compared to the observed proportions, examination of model parameters (along with their standard errors) can reveal places in the a priori hierarchy which are inadequately specified, thereby suggesting possible changes in the hierarchy. This is the important advantage offered by the Dayton and Macready model not offered by other available procedures.

The model does more than reject hypothesized hierarchies; it also suggests areas of the hierarchy where revision is necessary. Pushed to the limit, the procedure could be used as a pure discovery procedure from which a hierarchy could be built. This then would offer another method for building proposed hierarchies besides introspection and hierarchical clustering techniques. Dayton and Macready have mentioned that, at present, they have not utilized the procedure in this fashion. With the work of White, Dayton and Macready, and others (for example, Bart & Krus, 1973; Boozer & Lindvall, 1971; Macready, 1975), the statistical generation procedures and the necessary methods for generating and validating hierarchies now appear to be in a usable form.

Adaptive Testing Research

To date there have been only two investigations of adaptive testing to learning hierarchies (Ferguson, 1969; Spineti & Hambleton, 1977). The only other related study was by Vale (1977) and was an investigation of misclassification errors. Ferguson (1969) was concerned with classifying students as "masters" or "non-masters" on each objective in a learning hierarchy. The routing strategy was complex (involving the sequential ratio test described earlier) and required a computer to perform the actual routing. Ferguson found a 60% saving in number of items administered in the computerized administration using a variety of adaptive testing procedures. A test-retest of the adaptive testing procedure gave high reliability, with the reliabilities of the adaptive testing classifications higher than those of the paper-and-pencil conventional test approach.

An important consideration in the work of Spineti and Hambleton (1977) has been that the adaptive testing strategies under investigation be implementable without the aid of computer terminals. Such a restriction clearly sets this work apart from that of Ferguson and most of the other research on adaptive testing, with the exception of the self-scoring flexilevel testing work of Lord (1971). The primary effect of the restriction is that it eliminates the possibility of using complex decision-making rules such as the one adopted by Ferguson (1969). The concern has been to study the effectiveness of a multitude of adaptive testing strategies that could be implemented in objectives-based programs without the aid of computers. Since few objectives-based instructional programs have access to computer terminals for testing, this restriction was imposed so that the results would be of maximum usefulness. A fixed number of items was required to assess mastery of each objective tested, items were scored right or wrong, and all items measuring a particular objective were assumed to have similar statistical properties. Examinee performance on the test items was assumed to be represented by the binomial test model (Lord & Novick, 1968).

The interactive effects of several factors (test length, cutting score, and starting point) on the accuracy of mastery classification decisions and the amount of testing time in adaptive testing schemes were investigated. Values of each factor were combined to generate a multitude of adaptive testing strategies for study with two learning hierarchies and three different distributions of true scores across the hierarchies. The study was conducted via computer simulation techniques. Therefore, there was no need to be concerned about problems raised earlier in this paper, i.e., the problems of developing and validating criterion-referenced tests and learning hierarchies.

Of the many learning hierarchies reported in the educational literature, two were selected for study. These were the learning hierarchies for hydrolysis of salts (Gagné, 1970) and addition-subtraction (Ferguson, 1969). The second one was selected so some of the results of this study could be compared with Ferguson's results. The two learning hierarchies are shown in Figures 1 and 2 and are referred to as Hierarchy A and B, respectively. It was found that it was possible to obtain an overall reduction of more than 50% in testing time by introducing an adaptive testing scheme. With Gagné's hierarchy (Hierarchy A), the adaptive testing strategies resulted, on the average, in an overall reduction of testing time of 59.2%. With Ferguson's hierarchy (Hierarchy B), there was a 53.2% reduction in testing time. It is likely that adaptive testing strategies with Hierarchy B were not quite as effective as with Hierarchy A because Hierarchy B had two terminal objectives, whereas Hierarchy A had only one. The difference highlighted the importance of the particular form of the learning hierarchy on the effectiveness of adaptive testing strategies.

The results of this study on the saving of testing time varied from 50% to 70% and compared favorably with the empirical results of Ferguson (1969). He reported a saving of testing time of 60% over conventional testing. The similarity of the results added validity to the appropriateness of the simulation procedures of the present study.

The reduction in testing time derived from the adaptive testing strategies was impressive; however, it would have meant little if the total number of errors of classification was substantially larger than with conventional testing. In fact, with Hierarchy A the adaptive testing strategies resulted in a slightly lower number of errors of classification than with conventional testing. The reverse was true with Hierarchy B; but, again, the differences were slight. These findings, along with the information on the comparisons

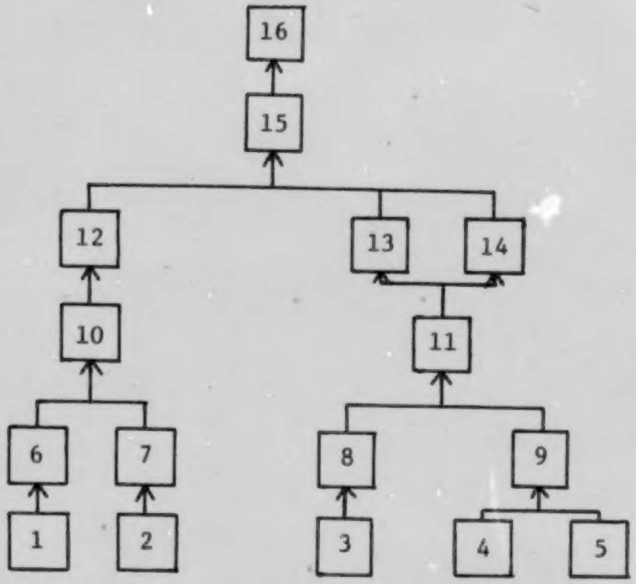

16
15
12
13
14
10
11
6
7
8
9
1
2
3
4
5

Figure 1. Gagne's Hydrolysis of Salts Hierarchy

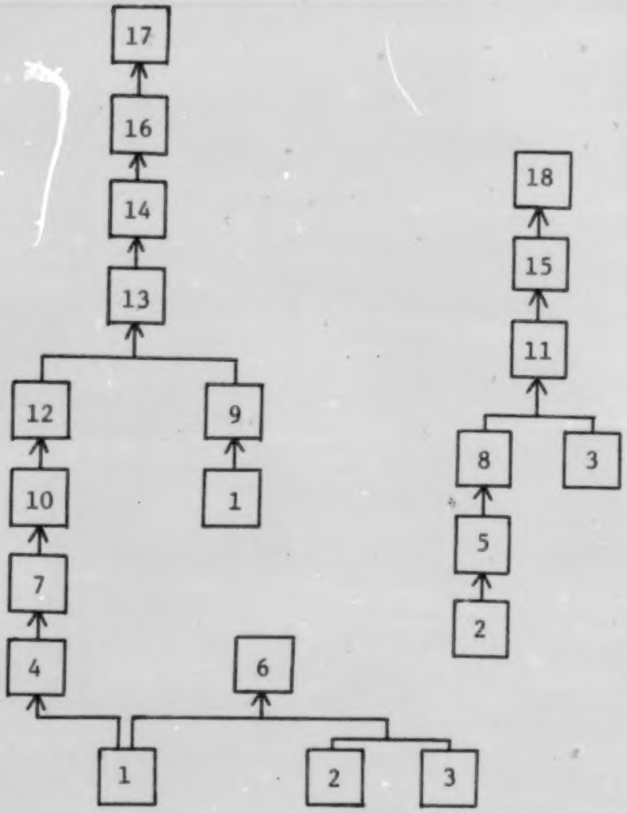

17
16
14
13
12
9
10
1
7
4
6
1
2
3
18
15
11
8
3
5
2

Figure 2. Ferguson's Addition-Subtraction Hierarchy

of testing time for conventional and adaptive testing, provide strong support for the speculations of many researchers. That is, by using an adaptive testing strategy in the context of learning hierarchies, there is much to be gained in terms of testing time efficiency without any significant loss in the accuracy of decision making.

It was dramatically clear from the numerous simulations that considerable saving in testing time was gained through implementing an adaptive testing strategy. Whereas the Ferguson adaptive testing strategies could only be implemented with the aid of computer-testing terminals, the Spineti-Hambleton testing strategies were simple enough to be implemented in a regular classroom with the aid of a "programmed instruction type" of test booklet.

Conclusions

The application of adaptive testing to learning hierarchies is substantially different from other applications and therefore includes some unique problems. First, while in any testing project there is concern about the generation of an item pool, the problem of developing an appropriate item pool for each criterion-referenced test is especially difficult; and the technology for accomplishing the task is quite new and not well understood. However, the "quality" of the item pool for a criterion-referenced test will be directly related to the overall success of an adaptive testing scheme. The problem must, therefore, be given careful attention.

Second, there is the unique problem of developing and validating learning hierarchies. Because of the inter-relationship between adaptive testing schemes and a learning hierarchy, the success of any adaptive testing scheme will depend on the "validity" of the learning hierarchy under investigation. In validating learning hierarchies,there are psychological as well as statistical problems involved. For example, several researchers have reported that while examinees may learn material in the sequence defined by a learning hierarchy, they may forget the information learned in any order. Thus, students may be able to perform a terminal objective although they have forgotten several of the prerequisite skills. The implications of this phenomenon for the validation of learning hierarchies and adaptive testing research are not clear. Third, classification problems,as opposed to measurement problems,are of interest. There has been relatively little research on using adaptive testing schemes to classify examinees into two or more categories.

There are adequate technologies to develop and to validate both criterion-referenced tests and learning hierarchies. Further refinements and advancements to the technology will take place as more researchers work in the area and encounter implementation problems. It will be especially interesting to observe the development of learning hierarchies and criterion-referenced tests in areas besides science and mathematics.

In adaptive testing research on learning hierarchies, it is important to distinguish between computer-assisted and non-computer-assisted test administration procedures. When computers are available for test administration, the possibility of using latent trait theory models and methods should be considered. To date, it is unclear how this application of latent trait theory could be accomplished. However, latent trait theory is now an established part of adaptive testing methodology for solving measurement problems. One possible problem may be that there usually are not sufficient items measuring any single objective to obtain a satisfactory latent trait ability estimate. This problem requires study as does the problem of optimal routing methods.

With non-computer-administered tests, some field studies need to be initiated. The design of such a study would involve developing a programmed instruction booklet which would include (1) test items designed to measure specific objectives in a learning hierarchy, (2) a self-scoring device, and (3) routing directions. Among the factors that could be investigated in an empirical study are test length, mastery cut-off score, and routing method. In addition, it would be interesting to study the merits, in terms of overall testing efficiency, of having individuals generate their own starting points for testing in the learning hierarchy.

References

Baker, R. B. Numerical taxonomy for educational researchers. Review of Educational Research, 1972, 42, 345-359.

Bart, W. M., & Krus, D. J. An ordering-theoretic method to determine hierarchies among items. Educational and Psychological Measurement, 1973, 33, 291-300.

Boozer, R. F., & Lindvall, C. M. An investigation of selected procedures for the development and evaluation of hierarchical curriculum structures. Pittsburgh, PA: University of Pittsburgh, Learning Research and Development Center, 1971.

Capie, W., & Jones, H. L. An assessment of hierarchy validation techniques. Journal of Research in Science Teaching, 1971, 8, 137-147.

Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, DC: American Council on Education, 1971.

Dayton, C. M., & Macready, G. B. A probabilistic model for validation of behavioral hierarchies. Psychometrika, 1976, 41, 189-204.

Ferguson, R. L. The development of a computer-assisted branched test for a program of individually prescribed instruction. Unpublished doctoral dissertation, University of Pittsburgh, 1969.

Gagné, R. M. The acquistion of knowledge. Psychological Review, 1962, 69, 355-365.

Gagné, R. M. Learning hierarchies. Educational Psychologist, 1968, 6, 1-9.

Gagné, R.M. The conditions of learning (2nd ed.). New York, NY: Holt, Rinehart, & Winston, 1970.

Gagné, R. M., & Bassler, O. C. Study of retention of some topics in non-metric geometry. Journal of Educational Psychology, 1963, 54, 123-131.

Gagné, R. M., & Paradise, N. E. Abilities and learning sets in knowledge acquisition. Psychological Monographs, 1961, 75, (14, Whole No. 518).

Glaser, R. Instructional technology and the measurement of learning outcomes. American Psychologist, 1963, 18, 519-521.

Glaser, R., & Nitko, A. J. Measurement in learning and instruction. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, DC: American Council on Education, 1971.

Hambleton, R. K. Testing and decision-making procedures for selected individualized instructional programs. Review of Educational Research, 1974, 44, 371-400.

Hambleton, R. K. Validation of criterion-referenced test score interpretations. A paper presented at the Third International Symposium on Educational Testing, University of Leiden, The Netherlands, 1977.

Hambleton, R. K., Hutten, L. R., & Swaminathan, H. A comparison of several methods for assessing student mastery in objectives-based instructional programs. Journal of Experimental Education, 1976, 45, 57-64.

Hambleton, R. K., & Novick, M. R. Toward an intergration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 47, in press.

Hambleton, R. K., Swaminathan, H., & Algina, J. Some contributions to the theory and practice of criterion-referenced testing. In D. N. M. de Gruijter & L. J. Th. van der Kamp (Eds.), Advances in Psychological and Educational Measurement. New York: Wiley, 1976.

Harris, C. W., Alkin, M. C., & Popham, W. J. (Eds.). Problems in criterion-referenced measurement (CSE Monograph Series in Evaluation, No. 3). Los Angeles, CA: University of California, Center for the Study of Evaluation, 1974.

Hively, W., Patterson, H. L., & Page, S. A. A "universe-defined" system of arithmetic achievement tests. Journal of Educational Measurement, 1968, 5, 275-290.

Johnson, S. C. Hierarchical clustering schemes. Psychometrika, 1967, 32, 241-254.

Linke, R. D. Replicative studies in hierarchical learning of graphical interpretation skills, *British Journal of Educational Psychology*, 1975, *45*, 39-46.

Linn, R. L. *Issues of validity in measurement for competency-based programs*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, 1977.

Lord, F. M. Some test theory for tailored testing. In W. H. Holtzman (Ed.), *Computer-assisted instruction, testing, and guidance*. New York, NY: Harper & Row, 1970.

Lord, F. M. The self-scoring flexilevel test. *Journal of Educational Measurement*, 1971, *8*, 147-151.

Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley, 1968.

Macready, G. B. The structure of domain hierarchies found within a domain-referenced testing system. *Educational and Psychological Measurement*, 1975, *35*, 583-598.

McQuitty, L. L. Hierarchical classification by multiple linkage. *Educational and Psychological Measurement*, 1970, *30*, 3-20.

Messick, S. A. The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, 1975, *30*, 955-966.

Millman, J. Passing scores and test lengths for domain-referenced tests. *Review of Educational Research*, 1973, *43*, 205-216.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), *Evaluation in education: Current practices*. San Francisco, CA: McCutchan, 1974.

Novick, M. R., & Jackson, P. H. *Statistical methods for educational and psychological research*. New York, NY: McGraw-Hill, 1974.

Passmore, D. L. *Sequencing learning events in performance-based instructional systems*. Paper presented at annual meeting of the Rocky Mountain Educational Research Association, Albuquerque, New Mexico, 1974.

Popham, W. J. *Educational evaluation*. Englewood Cliffs, NJ: Prentice-Hall, 1975.

Popham, W. J., & Husek, T. R. Implications of criterion-referenced measurement. *Journal of Educational Measurement*, 1969, *6*, 1-9.

Resnick, L. B., & Wang, M. C. Approaches to the validation of learning hierarchies. *Proceedings of the Eighteenth Annual Regional Conference on Testing Problems*. Princeton, NJ: Educational Testing Service, 1969.

Rovinelli, R. J., & Hambleton, R. K. On the use of content specialists in the assessment of criterion-referenced test item validity. Tijdschrift voor Onderwijsresearch, 1977, 2, 49-60.

Spineti, J. P., & Hambleton, R. K. A computer simulation study of tailored testing strategies for objectives-based instructional programs. Educational and Psychological Measurement, 1977, 37, 139-158.

Swaminathan, H., Hambleton, R. K., & Algina, J. Reliability of criterion-referenced tests: A decision-theoretic formulation. Journal of Educational Measurement, 1974, 11, 263-268.

Swaminathan, H., Hambleton, R. K., & Algina, J. A Bayesian decision-theoretic procedure for use with criterion-referenced tests. Journal of Educational Measurement, 1975, 12, 87-98.

Tyron, R. C., & Bailey, D. E. Cluster analysis. New York, NY: McGraw-Hill, 1970.

Vale, C. D. Adaptive testing and the problem of classification. In D. J. Weiss (Ed.), Applications of computerized adaptive testing (Research Report 77-1). Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1977. (NTIS No. ADA038114).

Wald, A. Sequential analysis. New York: Wiley, 1947.

Weiss, D. J. (Ed.). Applications of computerized adaptive testing (Research Report 77-1). Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1977. (NITS No. ADA038114).

White, R. T. Research into learning hierarchies. Review of Educational Research, 1973, 43, 361-375.

White, R. T. The validation of a learning hierarchy. American Educational Research Journal, 1974, 11, 121-136. (a)

White, R. T. A model for the validation of learning hierarchies. Journal of Research in Science Teaching, 1974, 11, 1-3. (b)

White, R. T. Indexes used in testing the validity of learning hierarchies. Journal of Research in Science Teaching, 1974, 11, 61-66. (c)

White, R. T., & Clark, R. M. A test of inclusion which allows for errors of measurement. Psychometrika, 1973, 38, 77-86.

Wood, R. Response-contingent testing. Review of Educational Research, 1973, 43, 529-544.

Acknowledgements

The project reported herein was performed pursuant to a grant from the National Institute of Education, Department of Health, Education and Welfare. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education and no official endorsement by the National Institute of Education should be inferred.

MULTI-CONTENT ADAPTIVE MEASUREMENT OF ACHIEVEMENT

DAVID J. WEISS
AND
JOEL M. BROWN
UNIVERSITY OF MINNESOTA

All the previous research in adaptive testing has been concerned with tests which covered only a single content area. Thus, all of the branching procedures implemented for adaptive selection of items to be administered to a testee have been designed exclusively for intra-test branching within a single, presumably unidimensional, content area. Unidimensional approaches to intra-test adaptive testing are useful for measurement in the achievement domain (e.g., Bejar, Weiss, & Gialluca, 1977; Bejar, Weiss, & Kingsbury, 1977). Frequently, however, achievement tests span several content areas. Consequently, in many cases the assumption of a single dimension may not be appropriate. For these kinds of achievement tests, or for achievement test batteries covering a number of separable content areas for which separate scores are required, none of the existing adaptive strategies (Weiss, 1974) are directly applicable.

There are two reasons why many of the adaptive testing strategies developed for single-content area ability tests may not be appropriate for achievement tests which cover several content areas. The first reason is that although the unidimensional branching models can be applied to separate content areas, they are not designed to take into account the information available between content areas. The second, and more practical, reason is that it might not be possible to generate relatively large numbers of items such as those required for many adaptive testing strategies within one content area in an achievement test. Urry (1977) has suggested that item pools to be used in adaptive testing with Owen's (1975) Bayesian testing strategy should include a minimum of 100 items to measure one dimension. Although there are no firm guidelines for other adaptive testing strategies, it is evident that they will function best with large item pools. Thus, application of these strategies to an achievement test battery of five subtests would require the test constructor to assemble 500 items with good psychometric qualities. Frequently, this is not possible. Consequently, in the application of adaptive testing to the unique problems in the measurement of achievement, an important research issue is the identification of adaptive testing strategies which make efficient use of existing item pools, rather than requiring the re-design of test item pools to meet the requirements of specific adaptive testing strategies.

The present paper describes an adaptive testing strategy which can be used in achievement tests with relatively small numbers of items. The strategy is designed for achievement test batteries or achievement tests with multiple content areas. It incorporates both intra-subtest branching and inter-subtest branching in order to efficiently adapt the test battery to each individual testee. The adaptive testing strategy is applied to a test battery and evaluated in terms of

1. The reduction in number of items administered,
2. Correlations of ability estimates with those derived from conventional administration of the test battery, and
3. The effects of adaptive administration on the psychometric information in the test scores.

METHOD

Purpose

The purpose of this study was to develop and evaluate an efficient and generalizable adaptive testing strategy for an achievement test battery comprised of a number of subtests. The adaptive testing strategy developed is designed to operate within a fixed item pool containing a relatively small number of items for each subtest. Real data simulation techniques (Weiss & Betz, 1973, pp. 11-12) were used. That is, the adaptive testing strategy was applied to item response data obtained from the administration of an achievement test battery which previously had been administered conventionally by paper-and-pencil. Results for the conventional testing strategy were compared with those for the adaptive testing strategy in terms of both test information and test length.

Procedure

Test Items and Subjects

Achievement test data were provided by the Personnel and Training Evaluation Program (PTEP) of the Naval Guided Missile School at Dam Neck, Virginia. These data were from a systems achievement test (SAT F17603) battery administered to 365 fire control technicians. The test battery included 12 subtests, each covering knowledge areas for different equipment or subject matter. Table 1 shows the content and number of items in each subtest. The test battery was administered in one booklet containing 232 items. The number of items per subtest ranged from 10 to 32; all of the items were multiple-choice with four response choices. The data provided by PTEP consisted of an identification number for each testee, the testee's number correct score on each of the 12 subtests, and correct-incorrect item responses for each of the 232 items.

Item Parameterization

Items were parameterized using Urry's ESTEM computer program for latent trait item parameterization employing the three-parameter normal ogive model (see Urry, 1976, p. 99). This program provided estimates of the item discrimination (*a*), item difficulty (*b*), and guessing (*c*) parameters. The items for each subtest were parameterized independently of items in other subtests.

Table 1
Number of Items in Each Subtest

Subtest	Content	No. of Items
A	Fire control system casualty procedures	10
B	Optical alignment group	10
C	Control console and power subsystem	18
D	Platform positioning equipment	22
E	Multiplexed equipment	18
F	Digital control computer and software	18
G	Digital control computer--operator interface	14
H	Magnetic disk file	12
I	Digital control computer--missile interface	24
J	Guidance and guidance testing	29
K	MTRE MKG MOD3	32
L	Spare guidance temperature monitor	25
Total		232

Adaptive Testing Strategy

The adaptive testing procedure was developed in order to reduce to a minimum the number of items administered to each individual with as little impact as possible upon the measurement characteristics of the test battery. Both intra-subtest adaptive branching and inter-subtest adaptive branching were used in the development of the procedure.

Intra-Subtest Branching

Item Selection. The basic concept for intra-subtest adaptive branching was that the order in which the items were to be administered was to be dependent upon values of the item information curve (Birnbaum, 1968, p. 462). For each item in each subtest, item information values were computed for values of θ ranging from -3.0 to +3.0 in steps of .2. Items were selected within a subtest for each testee by computing the value of all item information curves at the current estimated achievement level ($\hat{\theta}$) for that testee. The item selected for administration was the item which had the highest information value at the testee's current level of $\hat{\theta}$. Once an item was administered to a testee, it was eliminated from the subtest pool of available items for that testee.

Estimation of θ. Owen's (1975) Bayesian scoring procedure was used for this simulation study. This scoring procedure provides an achievement level estimate ($\hat{\theta}$) after each mth test item is administered. The procedure begins with a prior estimate of $\hat{\theta}_m$ and its variance (σ^2_m). For the first item of the first subtest administered (m=1), these were 0.00 and 1.00, respectively. An item was administered and scored as correct or incorrect. The revised estimate of $\hat{\theta}$ was determined using equations provided by Owen (1975, p. 353). The updated estimates of $\hat{\theta}$, along with their associated variances, were used as the prior estimates of $\hat{\theta}$ for the selection of the next test item, which was based on the maximum information rule described above. The next item was administered, and a new value of $\hat{\theta}$ was determined which was then used to select the next item. This procedure was repeated until a termination criterion was reached.

Termination criteria. Two criteria were used in determining when administration of items within a subtest should be stopped: (1) when all of the remaining items provided less than a pre-determined small amount of information or (2) when the within-subtest item pool was exhausted. Testing was terminated for a given testee at the first occurrence of one of these criteria within a given subtest. In applying the first criterion, testing was terminated when there was no item available which provided an information value greater than .01 at a given testee's current level of $\hat{\theta}$. Figure 1 diagramatically summarizes the intra-subtest branching procedure.

Figure 1
Intra-Subtest Branching Scheme

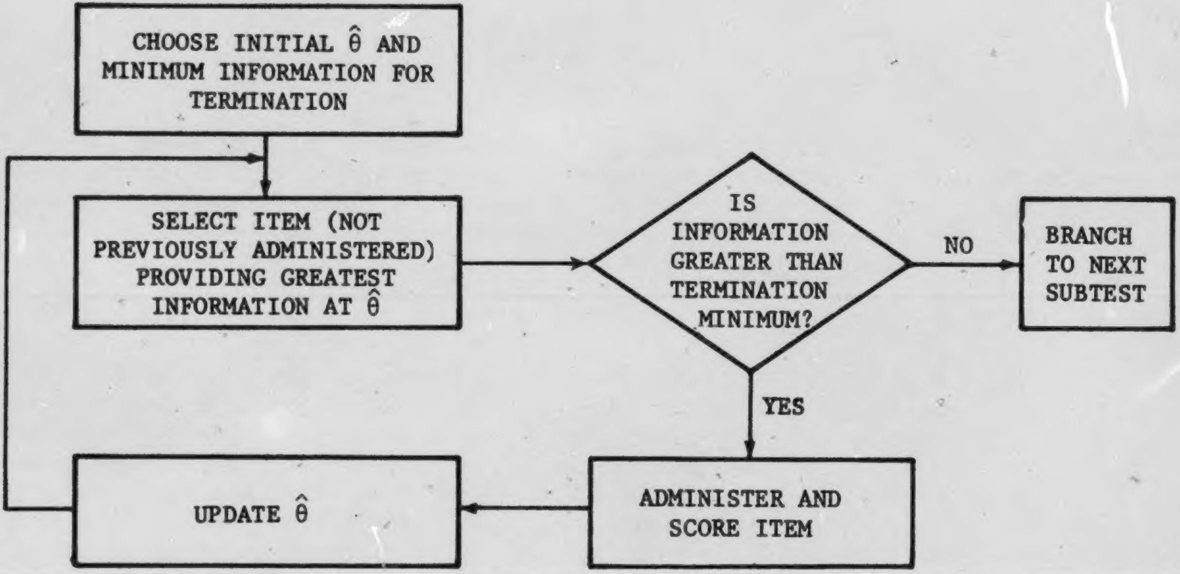
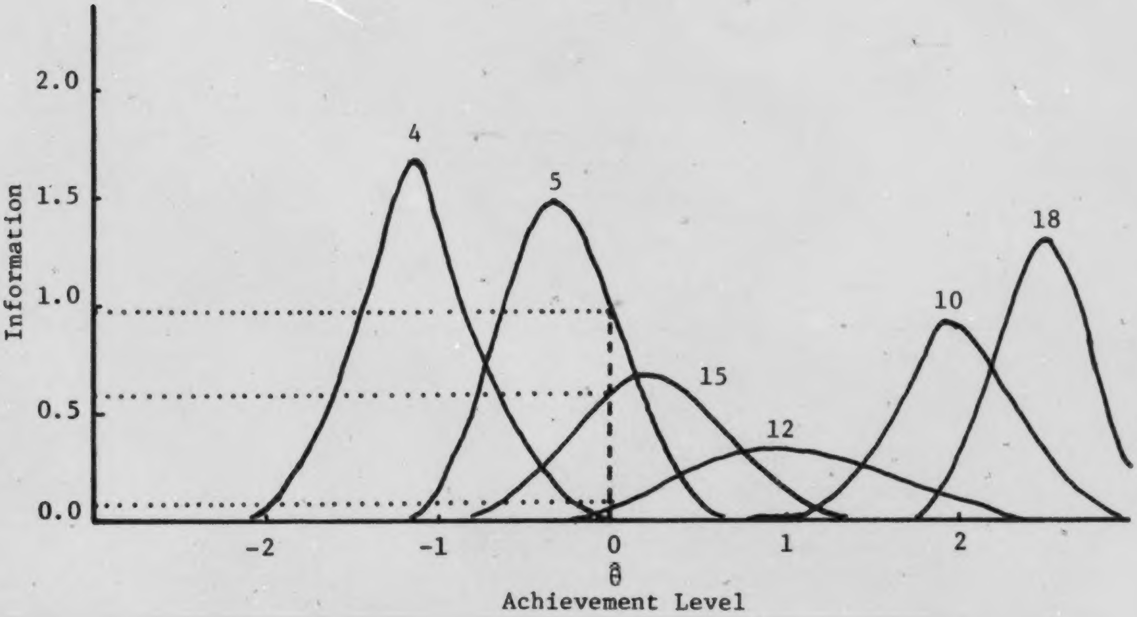

CHOOSE INITIAL $\hat{\theta}$ AND MINIMUM INFORMATION FOR TERMINATION
SELECT ITEM (NOT PREVIOUSLY ADMINISTERED) PROVIDING GREATEST INFORMATION AT $\hat{\theta}$
IS INFORMATION GREATER THAN TERMINATION MINIMUM?
NO
BRANCH TO NEXT SUBTEST
YES
ADMINISTER AND SCORE ITEM
UPDATE $\hat{\theta}$

Illustration of Intra-Subtest Adaptive Branching

Illustrating this procedure, Figure 2 shows estimated item information curves for 6 items from Subtest 1. (There were a total of 15 items in Subtest 1 from which only 6 were chosen to simplify the illustration.) The height of the information curve at a given achievement level indicates the amount of information provided by the item. Most of the items are fairly "peaked"; that is, they provide information over a relatively narrow range of the achievement continuum. While the information curves overlap to some degree, different items provide different amounts of information at a given point on the achievement continuum. The guiding principle for the adaptive procedure was to administer the item which provided the most information at the current achievement estimate.

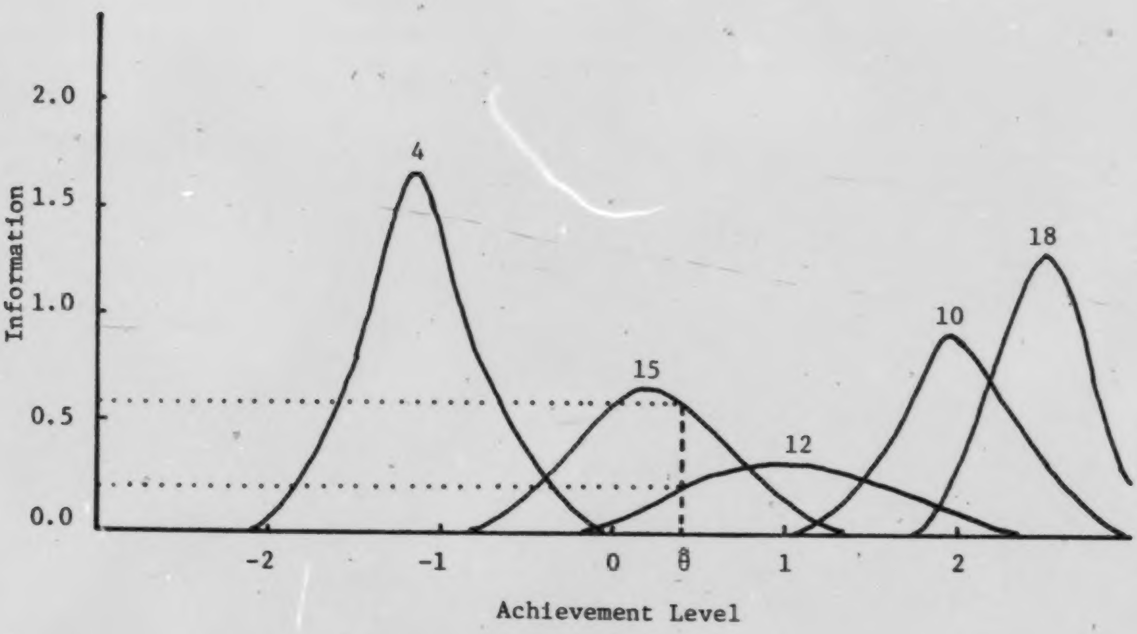
Figure 2
Estimated Item Information Curves for Six Items from Subtest 1



For a testee beginning Subtest 1, the initial achievement estimate was $\hat{\theta}=0$ (for subsequent subtests, this varied by individual); this is shown by the vertical dashed line in Figure 2. Of the six items in the example, only three items had essentially non-zero information values at $\hat{\theta}=0$. These values, shown by the horizontal dotted lines in Figure 2 were .90 for Item 5, .58 for Item 15, and .04 for Item 12. Applying the rule that the item selected is the one which provides the most information at the current $\hat{\theta}$, Item 5 would be selected for administration.

Figure 3 shows the revised value of $\hat{\theta}=.46$ derived from the Bayesian scoring routine, assuming that a correct answer was given to Item 5. The information curve for Item 5, which was already administered, is not shown in Figure 3. At the new value of $\hat{\theta}$, only Items 15 and 12 provide non-zero values of information. Since Item 15 has an information value of .54 and Item 12 has a value of .20, Item 15 is selected as the second item to be administered to this testee.

Figure 3
Estimated Item Information Curves for Five Items from Subtest 1



Assuming that the testee had correctly answered Item 15, the value of $\hat{\theta}$ increased to .92; this is shown in Figure 4. At that value of $\hat{\theta}$, Item 12 provides .32 information and Item 10 provides .02 information. Item 12 is thus administered next. Assuming that Item 12 was answered incorrectly, the $\hat{\theta}$ decreased to .62, which is plotted in Figure 5. The figure shows that of the three items remaining, none provides any information at the current level of $\hat{\theta}$. Thus, there is no need for administering additional items from Subtest 1; and testing in that subtest is terminated. The achievement level estimate of $\hat{\theta}_1$=.62 is taken as the testee's score on Subtest 1, since it is based on all items providing more than non-trivial amounts of information about that testee's achievement level.

Inter-Subtest Branching

Subtest ordering. The order of administration for the various subtests was chosen to take maximum advantage of the intercorrelations among them, thereby utilizing the redundant information in previously administered subtests. This was accomplished through linear multiple regression. First, the number correct subtest scores for the 12 subtests were intercorrelated; and the

Figure 4
Estimated Item Information Curves for Four Items from Test 1

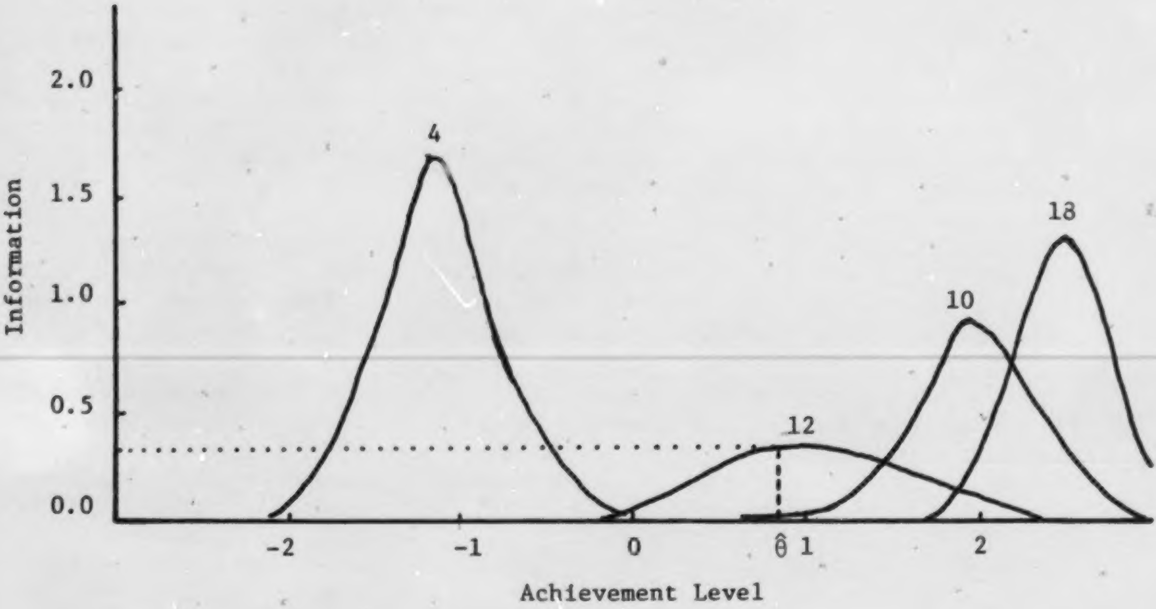
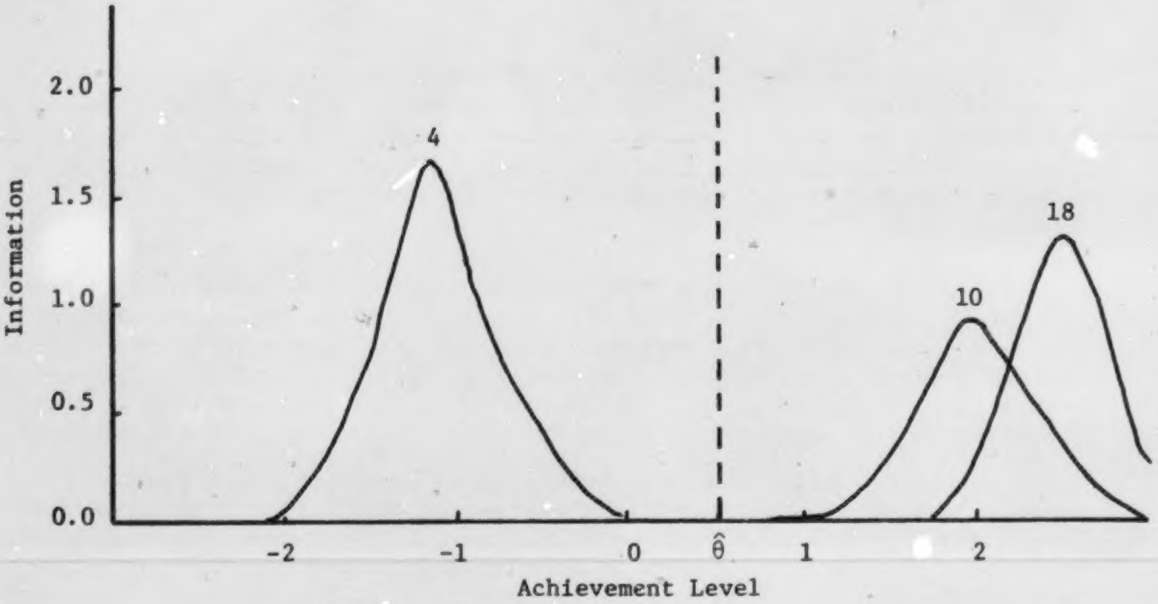


Figure 5
Estimated Item Information Curves for Three Items from Test 1



highest bivariate correlation was chosen from the intercorrelation matrix. One of these two subtests was arbitrarily designated to be administered first; the other was designated to be administered second.

Multiple correlations were then computed using the subtests previously designated first and second as predictor variables. Each of the 10 remaining subtests, in turn, was designated as the criterion variable. Of these 10 subtests, the one which had the highest multiple correlation with the first and second subtests was designated as the third subtest. This procedure was repeated to select the fourth subtest for the adaptive administration, computing multiple correlations with the first 3 subtests as predictor variables and each of the remaining 9 subtests, in turn, as the criterion variable. That subtest having the highest multiple correlation with the first 3 subtests was selected as the fourth subtest to be administered. By adding one subtest to the predictor set at each subsequent stage, this procedure was continued until all 12 subtests were ordered.

As a result of this procedure, the order in which the subtests were administered was the same for all testees. However, the selection of items within each subtest and the order in which those items were administered varied with testees as a function of the amount of item information provided at the testee's current achievement estimate.

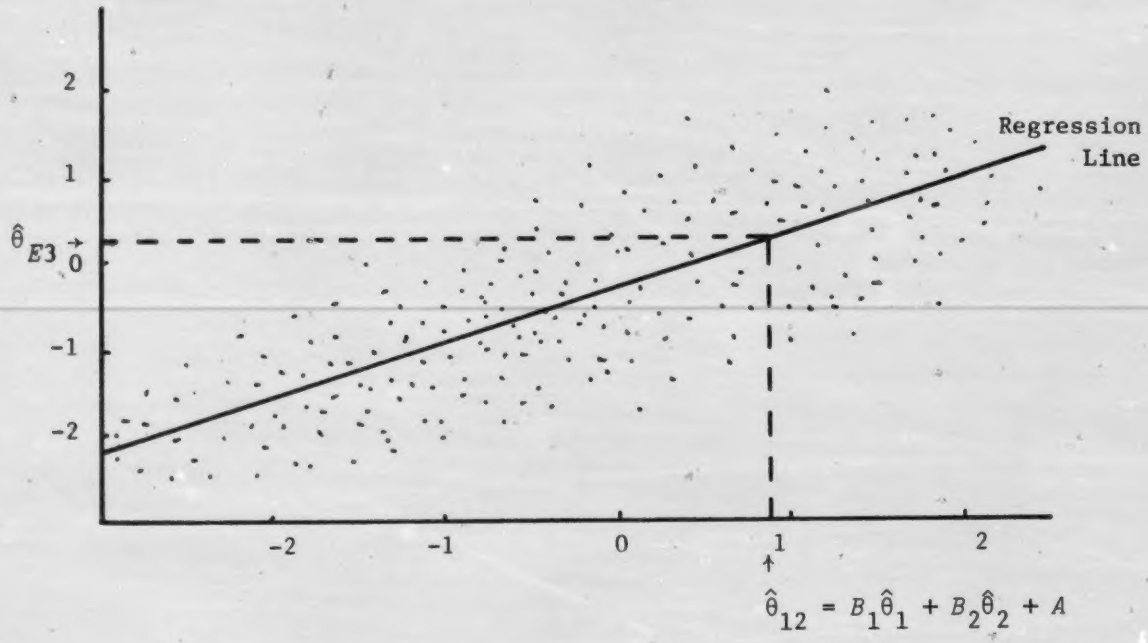
Differential subtest entry points. An important feature of the adaptive testing strategy implemented in this study was that after the first subtest, each testee's entry points for the second and subsequent subtests were differentially determined. For the first subtest, each testee's achievement level was assumed to be $\theta=0.00$. That is, having no previous information on which to base an estimate of the testee's achievement level, the initial item chosen from the first subtest for administration was the item which provided the most information for an estimated achievement level at the mean of the $\hat{\theta}$ distribution. Thus, all testees began the first subtest with the same test item.

The entry point into the item pool for the second subtest was determined from both the examinee's $\hat{\theta}$ at the end of the first subtest and the bivariate regression of scores from Subtest 1 on Subtest 2. This regression equation was based not only on scores for the items administered adaptively, but also on the correlations derived from number correct scores for all items in each of the subtests.

Determination of the entry point for the third and subsequent subtests was merely a generalization of the method used for the second subtest. The testee's achievement level estimates from Subtest 1 ($\hat{\theta}_1$) and Subtest 2 ($\hat{\theta}_2$) were entered into the multiple regression equation for predicting Subtest 3 scores from scores on Subtests 1 and 2. This generated an estimated subtest score for an individual ($\hat{\theta}_{E3}$) which was used as the initial prior achievement level estimate for intra-subtest branching in Subtest 3. The squared standard error of estimate from the multiple regression of Subtests 1 and 2 on Subtest 3 was used as the initial prior variance of the Bayesian achievement level estimate for Subtest 3. Figure 6 illustrates this differential entry point procedure.

Figure 6

Estimation of Initial Achievement Level Estimate for Subtest 3 ($\hat{\theta}_{E3}$) From the Multiple Regression of Subtest 1 ($\hat{\theta}_1$) and Subtest 2 ($\hat{\theta}_2$)



The inter-subtest branching regression procedure was used for entry into each of the remaining subtests. Each subsequent regression equation was based on the achievement estimates from each of the previously administered subtests. A testee's achievement level estimates for each subtest, based on the multiple regression of all previous subtests on a new subtest, was used as the initial Bayesian prior $\hat{\theta}$ for intra-subtest branching within that subtest. Item selection and scoring within subsequent subtests was then based on the intra-subtest branching procedures described earlier.

Conventional Test

A conventional test was used for comparison with the adaptive testing strategy. The subtests were administered in the same order for both the conventional and adaptive strategies. In the conventional strategy, all items within each subtest were administered sequentially so that all testees took the same items in the same order. Hence, there was no differential entry for the conventional strategy. In addition, all testees completed all items, which is typical in conventional testing. In order to facilitate comparison of results with the adaptive strategy, Bayesian scoring was employed for the conventional test. A mean of 0.0 and a variance of 1.0 were used as the initial prior achievement estimate of the Bayesian score for each subtest.

Data Analysis

The basic question examined in this study was whether the number of items administered could be reduced through adaptive testing without significantly changing the characteristics of the test scores. The effects of reducing the number of items by the adaptive testing item selection procedure were evaluated by means of both a correlational analysis and an information analysis.

Correlation Analysis

Early research comparing single test adaptive testing strategies with conventional testing strategies (see Betz & Weiss, 1973, 1974; Larkin & Weiss, 1974, 1975; Vale & Weiss, 1975; Weiss, 1973) demonstrated that adaptive tests resulted in test scores highly correlated with conventional test scores, even though the adaptive tests required substantially fewer items. Consequently, in the present study Pearson product-moment correlations were computed between subtest achievement level estimates ($\hat{\theta}$) from the conventional and adaptive testing procedures in order to examine the extent of the relationship between the scores. These were computed separately for each of the 12 subtests. High correlations between the scores would suggest that the tests ranked the examinees in a similar order along the achievement continuum.

Information Analysis

Information analyses were conducted in order to compare the adaptive and conventional testing strategies as a function of achievement levels. Test information values for different testing strategies at different levels on the achievement continuum provide an indication of their relative degree of precision of measurement (Birnbaum, 1968). Estimated test information curves were generated separately for each subtest for both conventional and adaptive testing strategies.

In the conventional testing strategy, an examinee's subtest information value was computed by summing the item information values at the examinee's final estimated achievement level ($\hat{\theta}$) for that subtest. An estimated information curve was plotted for the total group of examinees from their individual achievement level estimates and corresponding information values. For a conventional test, this is equivalent to computing the test information function using the item parameters, a, b, and c, as suggested by Birnbaum (1968, pp. 454-464).

Estimated subtest information curves were generated similarly for the adaptive testing strategy. The estimated value of test information was computed at each testee's final achievement estimate for the subtest by summing the information values at that $\hat{\theta}$ for the particular subset of items administered to that testee. Thus, for both adaptive and conventional testing, each test information value was computed at the final value of $\hat{\theta}$ for the subtest, based on the information provided by the items actually administered.

RESULTS

Test Length

The number of items administered under both the adaptive and conventional test strategies is summarized in Table 1. The data in Table 1 show substantial reductions in test length as a result of the adaptive testing strategy. For Subtest 1, 15 items were administered by the conventional procedure,while from 4 to 13 items were administered by the adaptive procedure. Fifty percent of the group answered between 7 and 10 items in the adaptive test. The mean number of items administered by the adaptive strategy in Subtest 1 was 8.73, which represents a 41.8% reduction from the number of items required by the conventional test.

Table 1
Number of Items Administered in 12 Adaptive and Conventional Subtests

Subtest	Conventional Test	Adaptive Test Mean	*S.D.*	Range Min	Max	Percent Reduction[a]
1	15	8.73	1.86	4	13	41.8
2	24	14.12	2.90	4	20	41.2
3	17	9.87	3.38	2	17	41.9
4	22	12.57	4.60	2	22	42.9
5	19	11.55	3.58	1	18	39.2
6	13	4.70	2.10	1	12	63.8
7	18	7.44	3.21	1	15	58.7
8	10	7.07	1.71	1	10	29.3
9	10	6.44	1.72	1	9	35.6
10	23	8.42	5.54	1	22	63.4
11	12	5.52	2.97	1	12	54.0
12	18	5.41	3.20	1	15	69.9
Mean	16.75	8.49	3.06	1.67	15.42	49.3
Test Battery	201	101.84	24.08	27	153	49.3

[a] Computed by the formula 100-[(Mean number of items in adaptive test/mean number of items in conventional test)×100]

Similar results were observed for the other subtests. Reduction of number of items required by the adaptive test varied from a low of 29.3% for Subtest 8 to a high of 69.9% for Subtest 12, in which a mean of 5.41 items was administered by the adaptive strategy. In Subtest 12, between 3 and 7 items were administered to 50% of the testees in the adaptive strategy as compared to 18 items for each testee in the conventional test. Subtest 12 had the highest percent reduction. In all probability, this was attributable to the increased accuracy of the test entry point from the multiple regression of the scores on the 11 prior subtests.

It is interesting to note that for Subtests 5 through 12, the minimum number of items administered by the adaptive procedure was one. For several of these subtests, a relatively substantial number of testees were administered only one item, i.e., almost 10% of the total group for Subtests 6, 11, and 12. The minimum number of items administered by the adaptive strategy was less for tests later in the adaptive testing sequence. This probably resulted from the increased use of prior test information for determining the initial item to be administered.

Although minimum numbers of items were administered at relatively high frequencies by the adaptive strategy, the maximum numbers of items were administered to very few testees. For Subtests 3, 4, 8, and 11 the maximum number of items administered by the adaptive strategy was the same as that administered by the conventional test; frequencies associated with these maximums were 2, 1, 5, and 1, respectively. For the remaining 8 subtests, none of the testees received the same number of items in the adaptive tests as they did in the conventional tests.

The conventional test battery consisted of 201 items administered to all testees. The average number of items administered by the adaptive strategy (see Table 1) was 101.84, representing a 49.3% reduction in number of items administered. The median number of items administered was 103, indicating a slight negative skew to the distribution. Fifty percent of the testees received between 86 and 119 items in the adaptive battery, representing reductions of 57.2% to 40.8% for half of the testees. None of the testees required all the items in the adaptive administration. The longest adaptive battery administered required 153 items for one testee, representing a 23.9% reduction in test length; the shortest adaptive battery for one testee required only 27 items, representing a test length reduction of 86.6%.

Correlation Analysis

Table 2 shows the Pearson product-moment correlations of the Bayesian achievement level estimates ($\hat{\theta}$) for the conventional and adaptive testing strategies: 11 of the 12 correlations were greater than .90. The highest correlations were .98 for Subtests 2 and 8; the lowest was .74 for Subtest 6.

Table 2
Correlation (r) of Bayesian Achievement Level Estimates ($\hat{\theta}$)
for the Adaptive and Conventional Testing Strategies by Subtest,
and Cronbach's Alpha Coefficient for the Conventional Subtests

Subtest	No. Items	r	Cronbach's Alpha
1	15	.91	.57
2	24	.98	.69
3	17	.96	.54
4	22	.97	.65
5	19	.93	.59
6	13	.74	.44
7	18	.90	.50
8	10	.98	.56
9	10	.95	.39
10	23	.92	.61
11	12	.91	.51
12	18	.94	.40

The items contributing to the Bayesian subtest achievement level estimates in the adaptive test were a subset of those used in the conventional test. Thus, to some extent, the magnitudes of the correlations in Table 2 were a function of this part-whole relationship. This is supported by a comparison with the Alpha internal consistency estimates for the conventional subtests shown in Table 2. If there were no part-whole relationship, the correlations between the achievement level estimates would be restricted by the internal consistencies. However, all the correlations were substantially higher than the Alpha values.

If the magnitude of the correlations of the two achievement estimates were primarily determined by the part-whole relationship attributable to common items, the number of items administered in a subtest would bear a strong relationship to these correlations. This was not generally the case: One of the two highest correlations (*r*=.98) was observed for Subtest 8, which had only 10 items in the conventional test, while Subtest 9, which also had 10 items, had an *r*=.95. Although Subtest 8 had the smallest percentage reduction attributable to the adaptive administration (20.3%; see Table 1), Subtest 9 had a 45.6% reduction and Subtest 2 (*r*=.98) had a 41.7% reduction. Subtest 6, which had the lowest *r* (.74), had a 63.8% reduction attributable to adaptive testing; but the highest percent reduction (69.9%) was observed for Subtest 12, for which an *r*=.94 was observed between the adaptive and conventional achievement estimates. Thus, these data suggest that the magnitudes of the correlations shown in Table 2 were not a direct function of either the number of items in the conventional tests or the internal consistency of those tests.

Information Analysis

Tables 3 and 4 provide mean raw values of estimated information [$I(\hat{\theta})$] at intervals of $\hat{\theta}$ for the adaptive and conventional tests for ordered Subtests 1 and 12. These values are based on mean information in test items actually administered to each testee, using the testee's $\hat{\theta}$ at the termination of each subtest.

Figure 7 shows a plot of estimated information values from adaptive and conventional administration of Subtest 1; estimated information values for the last subtest administered, Subtest 12, are shown in Figure 8. The information obtained from the adaptive administration of Subtest 1, for all practical purposes, was identical to the information from the conventional administration. The largest mean difference in information between adaptive and conventional administration, .14, occurred in the estimated achievement interval between $\hat{\theta}$=-1.39 and $\hat{\theta}$=-1.20. For Subtest 12 the information curves resulting from adaptive and conventional administration were practically identical except that the adaptive strategy produced a wider range of estimated achievement levels. That is, for Subtest 12 there were 46 testees who obtained adaptive scores which were less than the lowest of conventional scores.

CONCLUSIONS

This paper has presented an adaptive testing strategy designed for use with achievement test batteries covering multiple content areas. One goal of

Table 3
Adaptive and Conventional Test Mean Information Values [$I(\hat{\theta})$]
and Mean Difference in Information and t Values
at Estimated Achievement Levels ($\hat{\theta}$) for Subtest 1

$\hat{\theta}$ Interval		Adaptive Test			Conventional Test			Mean Difference		
Min	Max	N	$I_a(\hat{\theta})$	$S.D.$	N	$I_c(\hat{\theta})$	$S.D.$	$[I_c(\hat{\theta})-I_a(\hat{\theta})]$	t	df
-3.00	-2.80	0			0					
-2.79	-2.60	0			0					
-2.59	-2.40	0			0					
-2.39	-2.20	0			0					
-2.19	-2.00	0			0					
-1.99	-1.80	0			0					
-1.79	-1.60	11	.70	.29	14	.64	.23	-.06	-.58	23
-1.59	-1.40	22	1.85	.40	23	1.83	.35	-.02	-.18	43
-1.39	-1.20	21	2.87	.04	25	2.73	.18	-.14	-3.49**	44
-1.19	-1.00	23	2.86	.04	20	2.89	.03	.03	2.75**	41
-0.99	-0.80	25	2.86	.06	28	2.89	.06	.03	1.82	51
-0.79	-0.60	33	3.36	.24	37	3.38	.19	.02	.39	68
-0.59	-0.40	21	4.15	.15	19	4.15	.16	.00	.00	38
-0.39	-0.20	31	4.21	.11	24	4.26	.06	.05	2.01	53
-0.19	0.00	27	3.72	.19	32	3.75	.23	.03	.54	57
0.01	0.20	35	3.02	.21	30	3.04	.21	.02	.38	63
0.21	0.40	26	2.43	.09	31	2.50	.12	.07	2.45*	55
0.41	0.60	42	2.17	.04	29	2.23	.08	.06	4.17**	69
0.61	0.80	14	1.90	.00	27	1.96	.07	.06	3.19**	39
0.81	1.00	10	1.85	.00	10	1.81	.04	-.04	-3.16**	18
1.01	1.20	13	1.74	.00	7	1.74	.01	.00		
1.21	1.40	0			6	1.85	.01			
1.41	1.60	11	2.10	.00	3	2.13	.00	-.06		
1.61	1.80				0					
1.81	2.00				0					
2.01	2.20				0					
2.21	2.40				0					
2.41	2.60				0					
2.61	2.80				0					
2.81	3.00				0					

$^{*}p \leq .05$
$^{**}p \leq .01$

the strategy was to select and administer items within a subtest as a function of the amount of information provided by each item at each testee's current estimated achievement level. A second goal was to use redundant information between and among subtests (by predicting a testee's performance on subsequent subtests based on performance on previous subtests) to determine appropriate differential entry points in adaptive branching between subtests. It was hypothesized that attaining these goals in the design of an adaptive testing strategy would result in considerable reduction in the number of items

Table 4
Adaptive and Conventional Test Mean Information Values [$I(\hat{\theta})$]
and Mean Difference in Information and t Values
at Estimated Achievement Levels ($\hat{\theta}$) for Subtest 12

$\hat{\theta}$ Interval		Adaptive Test			Conventional Test			Mean Difference		
Min	Max	N	$I_a(\hat{\theta})$	$S.D.$	N	$I_c(\hat{\theta})$	$S.D.$	[$I_c(\hat{\theta})-I_a(\hat{\theta})$]	t	df
-3.00	-2.80	0			0					
-2.79	-2.60	0			0					
-2.59	-2.40	11	.11	.32	0					
-2.39	-2.20	7	.04	.01	0					
-2.19	-2.00	15	.06	.03	0					
-1.99	-1.80	13	.20	.04	0					
-1.79	-1.60	12	.41	.07	1	.53				
-1.59	-1.40	15	.88	.28	10	.95	.17	.07	.71	23
-1.39	-1.20	23	1.73	.24	21	1.81	.26	.08	1.06	42
-1.19	-1.00	23	2.63	.67	24	2.81	.32	.18	1.18	45
-0.99	-0.80	17	4.04	.24	31	3.86	.28	-.18	-2.24*	46
-0.79	-0.60	27	4.57	.14	32	4.56	.12	-.01	-.30	57
-0.59	-0.40	33	4.64	.83	44	4.80	.01	.16	1.28	75
-0.39	-0.20	23	4.75	.02	35	4.76	.02	[illegible]	1.86	56
-0.19	0.00	49	2.07	2.37	80	2.03	2.37	-.0[illegible]	-.09	127
0.01	0.20	19	4.97	.09	24	4.36	1.69	-.61	-1.57	41
0.21	0.40	16	5.23	.05	16	5.23	.06	.00	.00	30
0.41	0.60	10	5.25	.05	12	5.27	.04	.02	1.04	20
0.61	0.80	11	4.84	.17	10	4.89	.15	.05	.71	19
0.81	1.00	10	3.35	1.77	9	4.29	.16	.94		
1.01	1.20	4	3.73	.04	7	3.60	.09	-.13		
1.21	1.40	7	2.89	1.28	5	3.36	.05	.47		
1.41	1.60	4	3.30	.00	1	3.32		.02		
1.61	1.80	1	3.32		0					
1.81	2.00	1	3.69		0					
2.01	2.20	0			0					
2.21	2.40	0			1	6.66				
2.41	2.60	1	6.67		0					
2.61	2.80	0			0					
2.81	3.00	0			1	2.26				

* $p<.05$

administered to each testee, while sacrificing little, if any, test information compared to that obtainable by administering the entire test battery conventionally. Thus, the focus of this adaptive testing strategy was utilization of an existing item pool for an achievement test battery to efficiently measure or estimate each testee's achievement level.

The adaptive testing strategy described in this report (see Brown & Weiss, 1977, for further analyses of these data) provides methods for intra-subtest and inter-subtest branching which exclude the administration of unnecessary items. The data indicate that on this achievement test battery the length of the battery can be reduced by 50% for the typical testee. In no case was it

Figure 7
Observed Information Curves for Subtest 1 (Conventional
Test = 15 Items; Adaptive Test = 8.74 Items Average, Range = 4 to 13 Items)

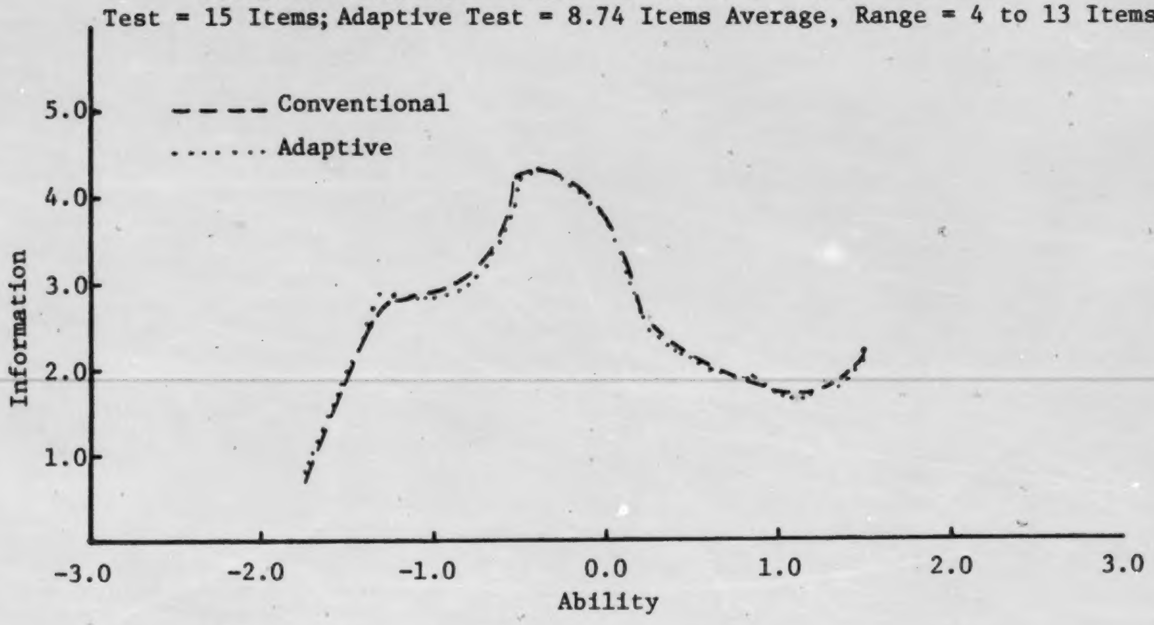
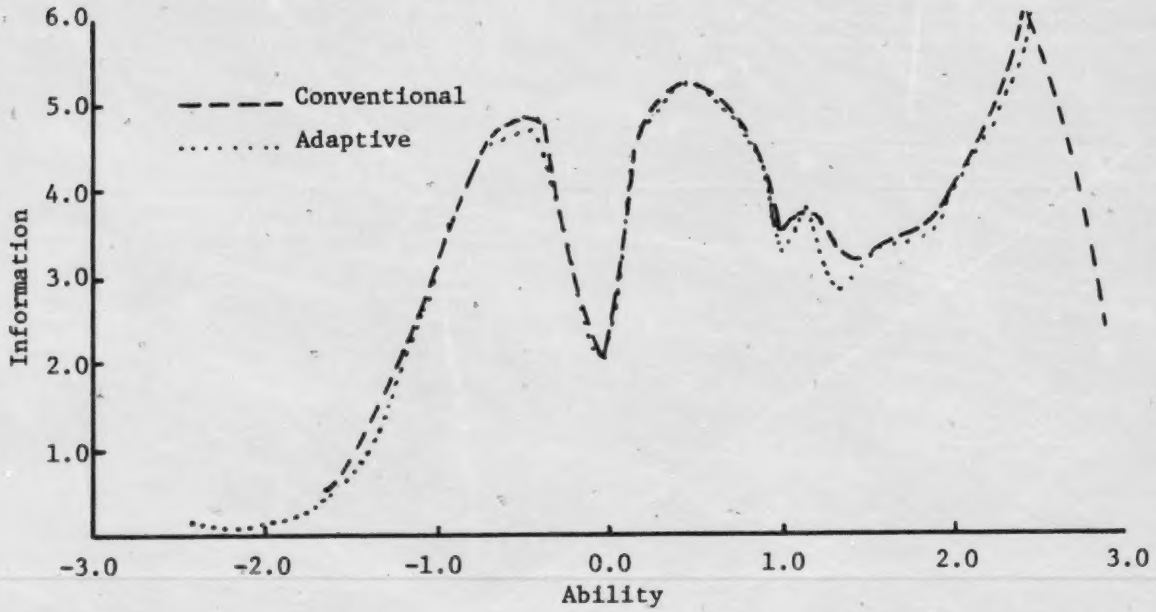


Figure 8
Observed Information Curves for Subtest 12 (Conventional
Test = 18 Items; Adaptive Test = 6.23 Items Average, Range = 1 to 15 Items)



necessary to administer in the adaptive battery all of the items included in the conventional tests. Scores from the adaptive tests correlated highly with those from the conventional tests. Adaptive testing therefore, can reduce the time spent in testing; the time saved could then be used by the testees for other activities, such as additional instruction. It is also possible that adaptive achievement testing might have positive psychological advantages (e.g., Betz & Weiss, 1976), providing further beneficial effects on the psychometric characteristics of test scores. At the least, reduced testing time might result in more favorable attitudes of the testees toward the testing process.

In the adaptive testing strategy implemented in this study, test length is a direct function of the termination criterion employed. Testing was terminated within a subtest when none of the remaining items had a corresponding level of item information greater than .01 (an arbitrarily chosen value) at the testee's current estimated achievement level. More research is needed to determine optimal termination criteria.

The results of this study have also shown that the amount of information extracted by adaptive testing closely approximated that for conventional testing. That the information curves resulting from the adaptive and conventional strategies were found to be highly correspondent was to be expected from the way in which items were selected (based on item information) for the adaptive strategy. However, because of the inapplicability of maximum likelihood scoring in the early stages of item administration within a subtest, additional research is needed to develop and evaluate optimal procedures for item scoring. In addition, further research is needed for identification and evaluation of optimal procedures to order subtests for inter-subtest branching.

This study has demonstrated that an adaptive testing strategy, designed specifically for achievement test batteries, can substantially reduce the number of items administered in all subtests of the battery without reducing the precision of subtest scores. The strategy appears to be generalizable; it should be applicable to a variety of test batteries in which there is a fixed and relatively small subset of items for each subtest. Further research is needed to evaluate the performance of this adaptive testing strategy in other test batteries and in live testing situations. In addition, research is needed to modify the adaptive testing strategy to identify optimal procedures for the complete individualized administration of an achievement test battery.

References

Bejar, I. I., Weiss, D. J., & Gialluca, K. A. An information comparison of conventional and adaptive tests in the measurement of classroom achievement (Research Report 77-7). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1977.

Bejar, I. I., Weiss, D. J., & Kingsbury, G. G. *Calibration of an item pool for the adaptive measurement of achievement* (Research Report 77-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1977.

Betz, N. E., & Weiss, D. J. *An empirical study of computer-administered two-stage ability testing* (Research Report 73-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1973. (NTIS No. AD 768993)

Betz, N. E., & Weiss, D. J. *Empirical and simulation studies of flexilevel ability testing* (Research Report 75-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1975. (NTIS No. AD A013185)

Betz, N. E., & Weiss, D. J. *Psychological effects of immediate knowledge of results and adaptive ability testing* (Research Report 76-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1976. (NTIS No. AD A027170)

Betz, N. E., & Weiss, D. J. *Simulation studies of two-stage ability testing* (Research Report 74-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1974. (NTIS No. AD A001230)

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley, 1968, chap. 17-20.

Brown, J. M., & Weiss, D. J. *An adaptive testing strategy for achievement test batteries* (Research Report 77-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1977.

Gugel, J. F., Schmidt, F. L., & Urry, V. W. Effectiveness of the ancillary correction procedure. In C. L. Clark (Ed.), *Proceedings of the first conference on computerized adaptive testing*. Washington, DC: U.S. Civil Service Commission, 1976.

Larkin, K. C., & Weiss, D. J. *An empirical comparison of two-stage and pyramidal adaptive ability testing* (Research Report 75-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1975. (NTIS No. AD A006733)

Larkin, K. C., & Weiss, D. J. *An empirical investigation of computer-administered pyramidal ability testing* (Research Report 74-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1974. (NTIS No. AD 783553)

Lord, F. M. Discussion. In W. A. Gorham, J. F. Gugel, F. M. Lord, C. Jensema, F. L. Schmidt, & V. W. Urry. *Computers and testing: Steps toward the inevitable conquest*. Symposium presented at the 83rd Annual Convention of the American Psychological Association, Chicago, 1975. (PB-261-694) Washington, DC: U.S. Civil Service Commission. (NTIS No. AD 16851469).

Owen, R. J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of the American Statistical Association, 1975, 70, 351-356.

Urry, V. W. A five-year quest: Is computer-assisted testing feasible? In C. L. Clark (Ed.). Proceedings of the first conference on computerized adaptive testing. Washington, DC: U.S. Civil Service Commission, 1976.

Urry, V. W. Tailored testing: A successful application of latent trait theory. Journal of Educational Measurement, 1977, 14, 181-196.

Vale, C. D., & Weiss, D. J. A simulation study of stradaptive ability testing (Research Report 75-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1975. (NTIS No. AD A020961)

Weiss, D. J. The stratified adaptive computerized ability test (Research Report 73-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1973. (NTIS No. AD 768376)

Weiss, D. J. Strategies of adaptive ability measurement (Research Report 74-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1974. (NTIS No. AD A004270)

Weiss, D. J., & Betz, N. E. Ability measurement: Conventional or adaptive? (Research Report 73-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1973. (NTIS No. AD 757788)

Acknowledgements

This research was supported by funds from the Army Research Institute, Air Force Human Resources Laboratory, Defense Advanced Research Projects Agency, Navy Personnel Research and Development Center and Office of Naval Research, and monitored by the Office of Naval Research under contract No. N00014-76-C-0627 NR150-389. Data utilized in this report were obtained from the Personnel and Training Evaluation Program (PTEP) of the Naval Guided Missile School at Dam Neck, Virginia. Appreciation is extended to Lieutenant Commander Lee J. Walker of PTEP and Dr. Myron A. Robinson of Data Design Laboratories at Norfolk, Virginia for their cooperation in this research.

Editor's Note

Although this paper was presented at the conference during this session, it was not on the final conference schedule. Since it was not available to the session discussant prior to the conference, discussion of the paper was not included in the discussant's comments.

DISCUSSION: SESSION 6

RICHARD L. FERGUSON
AMERICAN COLLEGE TESTING PROGRAM

A positive feature of the papers by Epstein and Knerr, Pennell, and Hambleton and Eignor is that they all addressed the topic of adaptive testing from the perspective of its applicability in an instructional context. That is, the papers examined adaptive testing in terms of its potential use in *applied* educational settings. Of particular interest to the authors were the promises of increased efficiency and accuracy of measurement via adaptive testing.

While their interest in applying adaptive assessment strategies in an instructional context is commendable, the authors tended to oversimplify or ignore some of the very germane considerations which ultimately will determine the viability of adaptive testing in that context. Believing these considerations to be almost as important as the methodology explored in the papers, I will comment on the former before addressing the latter.

Nearly all of the issues addressed by the three papers are significant and relevant from a psychometric perspective. Indeed, efforts to improve on the accuracy and efficiency of measurement methods are both worthwhile and laudatory. At the same time, many of the questions pursued in the papers, even when answered affirmatively, may prove to be of little consequence to educators. This contradiction can possibly be explained by what I believe to be the authors' somewhat inaccurate perception of how instruction proceeds in a classroom setting where adaptive testing might be implemented. To elaborate on this point, I will draw heavily on my own experience with adaptive testing in an applied setting--specifically, my earlier work with the Individually Prescribed Instruction Project (IPI) at the University of Pittsburgh (Ferguson, 1971). I will briefly describe the study which I conducted and then relate some of the ambiguities that surfaced with respect to the benefits of adaptive testing in that particular educational setting.

The adaptive testing study which I conducted involved the evaluation of students' levels of proficiency in each of 18 mathematics objectives/skills that were organized into a testing/learning hierarchy using a Gagné type task analysis. (The actual testing/learning hierarchy used in the study is presented as Figure 2 in the paper by Hambleton and Eignor.) The hierarchy was subsequently validated empirically by a process that included administration of tests measuring the 18 skills in sequence to about 200 students. Although a perfectly valid hierarchy was not achieved, sufficient data were generated to support the use of that hierarchy in an adaptive testing context.

There were two levels of adaptation involved in the IPI testing situation. It is crucial to distinguish between the two levels because each depended on a different decision theoretic and also led to a different conclusion about the merits of adaptive testing. The first level involved decisions about the particular objective/skill to be tested first and the order in which subsequent objectives/skills were tested. The second level involved decisions about what constituted sufficient proficiency in a particular objective/skill.

These two different levels of decision making occurred sequentially in my earlier study. For example, initial assessment using the computer-based approach to testing began with Objective 12 (an objective/skill in the middle of the hierarchy). For that objective/skill, the Wald sequential probability ratio test was then applied as items were administered and scored to determine whether or not the student had sufficient proficiency in the objective/skill. The cutoff score for sufficient proficiency and the acceptable classification error rates were set in advance; then the computer randomly generated items sampling the skill defined by Objective 12.

This particular approach was facilitated, of course, by the fact that the area tested was mathematics. No two students took precisely the same test, even if the routing was identical, because the computer itself was randomly generating the items after each decision analysis. After a mastery decision was reached for Objective 12, the examinee was branched for testing on another objective. After each mastery decision, branching continued until a mastery decision was made about all of the objectives/skills. In some cases, the mastery decision for an objective was made without any testing by using the prerequisite relationships implicit in the hierarchy. Branching from one objective/skill to another was determined as a function of the students' level of proficiency on the last objective for which a mastery decision had been reached and the previous mastery decisions about other objectives.

These adaptive procedures have been applied not only in the Ferguson (1969) study but also on a continuing basis in an actual school context as a regular part of the instructional testing program (Carlson & Fitzhugh, 1974). Considerable data relating to both the efficacy and viability of adaptive testing in instructional settings and the problems that occur in those settings were collected through these studies nearly a decade ago. For example, the Ferguson (1969) study shed some light on how mastery decisions about educational objectives/skills can be made with tolerable risks of classification errors using a Wald sequential probability ratio test. It also provided information about the merit of various strategies for branching from objective to objective in a testing/learning hierarchy during the testing. Use of the Wald test to ascertain students' proficiency in a particular objective/skill yielded almost no savings of testing time over the conventional testing. However, considerable savings in testing time occurred as a result of the particular strategy used to branch from objective to objective. In the latter case, adaptive testing resulted in about a 50 to 60% time savings over conventional testing.

This brief sketch of the procedures used in the Ferguson study provides a background to some general comments about the three papers presented. It seems that nearly every one of the papers begins with an implicit assumption. The assumption is that saving testing time is a goal relevant to the instructional program, one worthy of active pursuit. Although such an assumption is intellectually appealing, it is not a self-evident truth. More to the point, it is a testable assumption to which proper attention has yet to be given by eudcators and test developers. To date, most of the adaptive testing studies conducted, including the three described today, have been simulation studies. Consequently, they may have failed to consider the very practical questions and circumstances that are faced by classroom teachers. I shall amplify my point by describing a few of the practical matters that arose when I implemented adaptive testing in the school setting described earlier.

The value attached to saving time by adaptive testing is very much a function of the use which can be made of that time in the instructional setting once it has been saved. There are at least two major contexts in which adaptive testing can be used in the educational setting. The first context involves the use of the test in what is essentially a global assessment of students' proficiency with respect to some relatively large body of knowledge and skills. For example, into this category of use would fall testing aimed at a summative assessment of a year's work and assessment aimed at measuring general educational development, e.g., the Scholastic Aptitude Test (Cleary, Linn, & Rock, 1968). In this context, adaptive testing promises a savings of time that may be very meaningful.

The second major context for the use of adaptive testing in education is related to assessment pertinent to individualized instruction programs, that is, programs where both the testing *and* the instruction are adapted to students' capabilities. In this context, the purpose of the assessment is to "place" students in the appropriate module of the instructional sequence. The authors of the three papers presented today seem to be focusing primarily on this particular context for adaptive testing. Moreover, the papers seem to place a high value on the efficiency with which adaptive testing can yield decisions about students' proficiency in specified knowledge and skills.

It is my judgment that the benefits of efficiency attributed to adaptive testing in making instructional "placement" decisions are somewhat overstated. One cannot or should not ignore the practical question of whether the time saved is or can be used effectively to further students' educational growth. Unless the instructional setting is individualized (still a relatively rare occurrence in American education), it is highly unlikely that adaptive testing will contribute much to increased or more effective learning. Even if the instructional setting is individualized, no study has yet demonstrated that the amount of time saved by adaptive testing would yield dividends in students' learning. This is not a moot issue, since the instructional system must be carefully designed to insure that the student takes advantage of the time saved by engaging in productive learning activities. Obviously, the value of reduced testing time equates to the quality of how the time saved is invested.

Adaptive testing in instructional settings may also have some drawbacks. For example, classroom teachers frequently evaluate students' answers to test questions in the interest of "diagnosing" students' problems. This is especially true in the individualized instruction context. Reducing the amount of testing, particularly for objectives/skills that students have not mastered, also reduces the opportunity for teachers to inspect test items with errors, which in turn reduces their ability to prescribe instruction specifically directed at teaching the unlearned skills.

Another feature of adaptive testing is that it facilitates more accurate decisions about students' levels of proficiency in the objectives/skills measured. This is perceived as highly desirable because it is assumed that a high degree of precision is necessary in the assessment process if learning is to be optimized. In fact, such precision may not be essential at all. That is, some slippage in the evaluation of students' proficiencies may be inevitable and not necessarily detrimental to students' development over the long term.

From a related perspective, students may actually benefit more from testing that samples a large portion (as opposed to a small portion) of the objective/skill being tested. Testing can serve to reinforce learning and indeed may be an essential component in the learning process. To reduce the amount of testing in the interest of saving time may prove counterproductive. This issue deserves more exploration than it has been given to date.

Another concern experienced firsthand in the IPI context involves the procedures that Hambleton and others have described, whereby hierarchies and tests are developed with great concern for accuracy. That process requires a considerable investment in time and resources. As a result, such procedures tend to create inflexibility in the instructional process. For example, in the IPI study great care was taken to develop valid hierarchies and adaptive tests. Obviously, this must be done for each unit of instruction if a branching strategy is to be used. One outcome of this process was to lock instruction into the structure used in the testing. Indeed, modifying the instructional program inevitably affected the hierarchy and the adaptive test related to it. Some changes in the hierarchy and the test could only be made at considerable expense and inconvenience. Neither of the two situations described above is very desirable.

Implicit in Epstein and Knerr's paper (and perhaps in the other papers as well) is the view that there is a single particular sequence appropriate for learning a set of instructional objectives/skills. This is simply another assumption that is not fully supported by the evidence. Although data can be collected that establish prerequisite relationships for many sets of objective/skills, not all instructional objectives/skills can be organized into hierarchies which indicate prerequisite relationships. As noted earlier, even when the objectives/skills can be organized into hierarchies, use of the relationships implicit in the hierarchy to reduce the amount of testing can be problematic. For example, many of the objectives identified in the hierarchy demonstrated in the Hambleton and Eignor paper might not be tested at all for individual students, depending upon their initial test performance. If they failed to master Objective 12, for example, they were not tested in any of the higher level objectives/skills. The information that would have been generated from the testing of those objectives, had they not been bypassed, probably would have influenced the instructional prescription for the student.

The Epstein and Knerr paper expressed the viewpoint that having to specify meaningful criteria in the process of implementing a model of the

Wald type will result in greater thought being given to the misclassification problem (incorrectly declaring a student a master or a nonmaster). I do not believe that we can count on this occurring, partly because the setting of these criteria essentially involves value judgments. We may also be well advised to avoid investing significant effort in seeking ways to further minimize classification errors in an instructional context. To seek such methods would probably be putting effort in the wrong direction.

It is highly questionable, in my opinion, whether great precision is needed in assessment used for instructional decisions. Most systems, whether they are individualized or not, tend to have checkpoints built into them that enable the teacher to correct for the misclassification or misjudgment about a student's proficiency. Our current system of education is a testimonial to this fact. This does not mean, however, that we should not employ available means for improving the accuracy of measurement.

The Wald approach does offer the tester a way of improving on the accuracy of decisions about students' level of functioning. The assumptions that underlie the use of the Wald test have been demonstrated to hold when it is used as described in the Ferguson study (1971). Indeed, a dissertation by Cotton (1971) at the University of Pittsburgh examined its robustness, and the results were quite supportive. Thus, the Wald approach deserves consideration when it can be implemented in a practical way.

The future of adaptive testing in an instructional setting is as positive today as it was 10 years ago. The possibilities and promises are real. I would, however, encourage researchers to place greater emphasis on empirical studies so that the questions asked about adaptive testing are more relevant to the purposes for which they are used.

References

Carlson, M. & Fitzhugh, R. A computer-assisted instructional system for elementary mathematics. Pittsburgh, PA: University of Pittsburgh, Learning Research and Development Center, 1974.

Cleary, T. A., Linn, R., & Rock, D. A. An exploratory study of programmed tests. Educational and Psychological Measurement, 1968, 28, 345-360.

Cotton, T. S. An empirical test of the binomial error model applied to criterion-referenced tests (Doctoral dissertation, University of Pittsburgh, 1971). Dissertation Abstracts International, 1972, 32, 6186A. (University Microfilms No. 72-16,062).

Ferguson, R. L. Computer assistance for individualizing measurement. Pittsburgh, PA: University of Pittsburgh, Learning Research and Development Center, 1971.

SESSION 7

ACHIEVEMENT TESTING VIEWED AS A TRAIT MEASUREMENT PROBLEM

APPLICATIONS OF LATENT TRAIT THEORY
TO CRITERION-REFERENCED TESTING

JAMES R. MCBRIDE
ARMY RESEARCH INSTITUTE

UNIFACTOR LATENT TRAIT MODELS
APPLIED TO MULTIFACTOR TESTS:
RESULTS AND IMPLICATIONS

MARK RECKASE
UNIVERSITY OF MISSOURI

A COMPARISON OF CONVENTIONAL AND
COMPUTER-BASED ADAPTIVE
ACHIEVEMENT TESTING

ISAAC I. BEJAR
UNIVERSITY OF MINNESOTA

DISCUSSION

HARIHARAN SWAMINATHAN
UNIVERSITY OF MASSACHUSETTS

SESSION 7: ABSTRACTS

APPLICATIONS OF LATENT TRAIT THEORY TO CRITERION-REFERENCED TESTING

JAMES R. MCBRIDE

Criterion-referenced testing is considered to subsume both mastery testing and diagnostic testing. This paper demonstrates that (1) latent trait theory embodies techniques useful for test design, analysis, and scoring in a criterion-referenced framework; (2) the combination of latent trait test theory, domain sampling, and adaptive testing can provide a highly useful methodology for accurate, efficient assessment of performance in a multi-faceted domain of performance tasks; and (3) the extension of latent trait theory to criterion-referenced measurement provides a single test theory applicable to both diagnostic evaluation and mastery classification.

UNIFACTOR LATENT TRAIT MODELS APPLIED TO MULTIFACTOR TESTS: RESULTS AND IMPLICATIONS

MARK D. RECKASE

One- and three-parameter logistic test models and multidimensional data reduction procedures were applied to three types of test data: (1) classroom test data; (2) standardized achievement test data; and (3) simulated test data of varying factorial complexity, with sample sizes varied to determine the utility of the two logistic models. Results showed that (1) the logistic models estimate the first principal component of a test, even when that component accounts for less than 10% of the variance of the test; (2) the one- and three-parameter models yielded essentially the same results on factorially complex tests when the sample sizes were small to moderate; and (3) factor analytic procedures can be used to sort items into unidimensional content components. The effect of multidimensionality on item calibration is also discussed. It was concluded that if a generalized achievement variable related to total score is of interest, latent trait models are directly applicable to existing achievement tests. If achievement estimates in specific content areas are desired, however, unidimensional subtests with separate item calibrations must be produced.

A COMPARISON OF CONVENTIONAL AND COMPUTER-BASED ADAPTIVE ACHIEVEMENT TESTING

ISAAC I. BEJAR

The stradaptive testing strategy was used to measure achievement in an introductory Biology course, and its efficiency was compared to a typical paper-and-pencil classroom examination. The Biology item pool was calibrated using item characteristic curve techniques; subsequent analyses indicated that the responses of a new group of students to these items were generally unidimensional. The validity of the adaptive and classroom tests were compared by fitting a factor model to the correlations among the adaptive and classroom tests at two points in time. The results suggested that, after taking test length into account, the adaptive test was superior to the classroom test; an information analysis corroborated these results. It is concluded that adaptive achievement testing is not only feasible but also an improvement over conventional testing.

APPLICATIONS OF LATENT TRAIT THEORY TO CRITERION-REFERENCED TESTING

JAMES R. MCBRIDE
U.S. ARMY RESEARCH INSTITUTE FOR THE
BEHAVIORAL AND SOCIAL SCIENCES

Considerable attention in recent years has been given to the distinctions between norm-referenced testing and criterion-referenced testing and to the lack of an adequate body of test theory for the latter. Some writers suggest that the distinction is largely artificial and that classical test theory applies to criterion-referenced or objective-referenced tests as a special case. An opposing view holds that the conceptual orientation of criterion-referenced testing makes it differ fundamentally from normative-based testing and that such differences, combined with the possible absence of variability in a group's raw test scores, make the psychometric tools of traditional test theory inappropriate for use in the design, analysis, and interpretation of results of criterion-referenced tests. The thesis presented in this paper is that the psychometric principles and procedures embodied by latent trait theory (also called item characteristic curve theory or item response theory) are broadly applicable to the problems of criterion-referenced, as well as norm-referenced, testing. In fact, they may provide the means for a substantial rapprochement between the two divergent schools of thought, typified by the sharpening distinction between the two purposes of testing.

The Nature of Criterion-Referenced Testing

The purpose of criterion-referenced tests is that they are intended to be used for performance evaluation. Such evaluation may be of individuals or groups and may be formative or summative; its essential feature is that it includes an objective, quantitative appraisal of what the person being evaluated can do at the time of evaluation, with respect to the performance domain of interest. This appraisal may be single-valued, as in the case of a unitary or homogeneous domain of performance tasks, or multi-valued, as in the case of a multi-faceted or heterogeneous domain. The appraisal may be a proficiency estimate (e.g., the percentage of domain tasks which the examinee can perform successfully); a mastery classification (e.g., determining for which of several discrete proficiency categories the examinee's current performance qualifies him/her); or a diagnostic assessment. The critical element is that it be "directly interpretable in terms of specified performance standards" (Glaser & Nitko, 1971).

Possibly the most general example of a directly interpretable criterion-referenced measure is the percentage score, which may be a straightforward summary of the examinee's performance on the specific items comprising the test or an estimate of domain "true score"--the proportion of tasks in the performance domain which the examinee can perform successfully. This "domain true score" seems to be an obvious and unobjectionable desideratum in criterion-referenced measurement, since it is a directly interpretable indicator of what a person can do in the domain of interest.

One objective of a test theory for criterion-referenced testing might be to provide an estimate of domain true score. Traditional norm-referenced test theory is inadequate for this purpose: A norm-referenced test score is not directly interpretable with respect to a domain of performance tasks. Furthermore, the item selection criteria typically used in the construction of norm-referenced tests may result in substantial bias in the content of a test intended to assess performance, so that performance on the test is not representative of performance in the task domain. These considerations led to Kriewall's (1972) specification of an item-sampling theory for criterion-referenced test construction and to the "domain-referenced" measurement prescriptions of Hively, Patterson, and Page (1968).

Kriewall (1972) developed the rudiments of a psychometric basis for criterion-referenced testing, which stopped somewhat short of an item response theoretic approach. He proposed constructing instructional tests by sampling items at random from an item population defined in terms of a specified learning objective--performance tasks which specify the intended outcomes of a segment of instruction. The individual's relative (proportion correct) score is interpreted as an estimate of proficiency with respect to the item population defined, where proficiency is defined as relative true score on the finite population of items measuring a given learning objective. Kriewall explicitly assumed that an individual's proficiency is fixed at a given point in time and that local independence holds.

In Kriewall's work, an item's p-value in the population of persons was irrelevant for criterion-referenced testing purposes. What was of interest was the conditional difficulty--probability of a correct response for a given level of proficiency. Although Kriewall acknowledged that conditional difficulty could vary from item to item, his psychometric development implicitly treated all items as having equal conditional difficulty. His formulation also treated all items as equi-discriminating, and he recommended that multiple-choice items not be used for proficiency estimation purposes.

Item Response Theory and the Performance Domain

A rationale for extending item response theory to criterion-referenced testing proceeds from a motivation similar to that underlying Kriewall's development: The purposes of criterion-referenced testing can be well served by a psychometric foundation which provides estimates of domain true score. Domain true score is herein defined as relative true score across a specified domain of performance tasks, i.e., as the proportion of tasks in the performance domain which the individual can accomplish successfully. Performance tasks may include test items (production and multiple-choice) as well as practical demonstrations of acquired knowledge and skills. A performance domain may be single- or multi-faceted; each facet may have one or more subdomains of content, and each content subdomain may consist of tasks with one or more levels of complexity. The definition of a domain is the specification of task classes constituting each facet, content subdomain, and level of complexity, and the relative importance of each.

To simplify exposition, consider a hierarchical taxonomy in which the performance domain consists of several task classes which, in turn, consist of task elements. Task classes may be somewhat arbitrarily defined. The task element is the fundamental building block in the construction (and definition) of the performance domain. An individual's proficiency in any domain is simply a composite of his/her proficiency in the task elements which constitute the domain.

Components of the Performance Domain

Task element. A task element is considered to be a unitary and explicitly defined skill on which proficiency can be measured using homogeneous test items or performance samples. Both *conceptual* and *response* homogeneity (Harris, 1974) are indicated here, interchangeable items having identical conditional probabilities of correct response. An individual is said to have a *true score* on any task element. This is defined as the probability of a correct response (or successful task performance) to an item (or task) sampled at random from the universe of items measuring the task element.

Alternatively, task element true score may be interpreted as the proportion of task element items an individual can answer correctly or tasks he/she can perform successfully. Let P_{ig} indicate this relative true score of person i on task element g. Conceptually, P_{ig} is very similar to the proficiency level measure used in defining *mastery*: Person i is a master if P_{ig} exceeds an arbitrary value, π_o. It is likewise similar to the construct measured in an individual formative evaluation test.

Task class. A task class is a composite of task elements, weighted arbitrarily as to their relative importance. Task class proficiency is defined as a weighted composite of task element proficiency values:

$$P_{ic} = \sum_{g=1}^{k} w_{gc} P_{ig} , \qquad [1]$$

$$0 < w_{gc} < 1$$

$$\sum_{g} w_{gc} = 1 ,$$

where k is the number of task elements comprising the task class c, and w_{gc} is the relative importance of element g in the task class.

A task class might consist of task elements related to a single content area but differing in complexity.

Task domain. A task domain is defined as a composite of task classes, again weighted proportionately to importance. Task domain proficiency, or domain true score, is the weighted composite of task class proficiency values:

$$P_{it} = \sum_{c=1}^{m} w_{ct} P_{ic} \quad , \qquad [2]$$

$$0 < w_{ct} < 1$$

$$\Sigma w_{ct} = 1 \quad .$$

Domain true score is thus a relative frequency weighted average of proficiency on several task classes or elements. It is similar to the variable being measured in a summative evaluation test.

Ultimately, task class and domain proficiency are defined in terms of either proficiency on task elements or the individual's conditional probability of successful task element performance. The matters of defining the domain structure and determining the relative importance of task classes (and task elements within each class) are substantive considerations appropriately left to subject matter experts. The problem of estimating the conditional probabilities for task elements is a psychometric problem to which item response theory can be applied.

Application of Item Response Theory

Under item response theory, an attribute, θ, is postulated to vary continuously and to be related to certain behaviors by a probabilistic function so that, for example, the conditional probability of correct response $P_j(\theta)$ to a particular test item, j, is a monotonic increasing function of the attribute. The function relating response probability to the attribute is the item characteristic function, which has specifiable form and parameters. For a given item or response, the parameters are assumed constant so that the probability varies solely as a function of the attribute θ. Individuals, i, are characterized by their locations, θ_i, on the attribute scale. An individual's scale location (attribute level) is assumed to be constant at a given point in time, so that if his/her location, θ_i, and the form and parameters of an item characteristic function are known, the probability of a correct response can be calculated exactly. This probability can be interpreted as (1) the proportion of correct responses to a specific item in the population of persons at attribute level θ_i or (2) the expected proportion of correct responses by an individual, i, in an infinite population of items having the identical characteristic function. It is the latter interpretation which will be exploited for the purposes of this paper.

Recall that a task element was said to be an explicitly defined skill measureable by a set of homogeneous items (or behavior samples) having identical conditional probabilities of correct response. Let the attribute θ be achievement level in the performance domain. It is assumed that achievement varies continuously, that an individual's achievement is fixed at any given point in time, and that intra-individual achievement varies as a function of time (or experience, or instruction). Performance on any item or task which measures a task element is assumed to be related stochastically to achievement by a monotone increasing function, the item response function

$$P_j(\theta) = F(\theta, \beta_j) \quad , \qquad [3]$$

where $P_j(\theta)$ is the probability of a successful performance on item j, conditional on achievement level;
$F(\ )$ is a generalized monotone function; and
β is a set of parameters specifying item j's response function.
Assume that the functional form of the response function is determined by the scaling chosen for θ and that the item parameters are fixed.

If the forms and parameters of a set of k performance items are known, the number-correct true score can be calculated exactly for any value of θ (Lord, 1977):

$$T(\theta) = \sum_{j=1}^{k} P_j(\theta) \quad . \qquad [4]$$

The relative true score is simply $T(\theta)/k$. For a given task element, appropriate items were said to have identical conditional probabilities; hence, they must have identical response functions. Therefore, the task element proficiency score for any individual, i, is exactly equal to the local value of the item response function of any item measuring that task element:

$$P_{ig} = E\,[P_j\,(\theta_i)] = P_j(\theta_i) \ , \text{ all } j \in g \quad . \qquad [5]$$

An example in the arithmetic domain. Figure 1 can be considered as an illustration of this conceptualization. A substantive performance domain, Arithmetic Operations, is cross-classified as to type of operation (i.e., addition, subtraction) and level of complexity of operands (say, 1-, 2-, and 3-digit numbers). Each row or column depicted may be considered a task class. Each cell in the matrix defined by the rows and columns is a task element, the items measuring which are homogeneous as to content and difficulty within any level of achievement. The number in each task element cell is the (hypothesized) conditional probability of successful performance on any appropriate production-type performance item for a given level of achievement, e.g., the ninth month of sixth grade. The curriculum is not geared, however, to performance at the rather microscopic task element level; instead, students are evaluated at a higher level as to their competence on classes of task elements. Thus, in the example, a student's proficiency in addition is a composite of his/her competence for addition operations involving 1-, 2-, and 3-digit operands.

Figure 1
Conditional Task Element Proficiency Values for the Arithmetic Operations Performance Domain

Operation	Complexity Low	Complexity Middle	Complexity High
Addition	.99	.90	.80
Subtraction	.95	.85	.75
Multiplication	.90	.80	.70
Division	.85	.75	.65

In constructing a mastery test of addition operations, the relative importance of the three levels of complexity would probably be reflected in the proportion of test items chosen at each level. Analogously, if the conditional probabilities in the first row of Figure 1 were known, they could be weighted in the same proportions into a composite task class proficiency score, such as

$$\begin{aligned} P_{ic} &= .25\ P_{11}(\theta_i) + .25\ P_{12}(\theta_i) + .50\ P_{13}(\theta_i) \qquad [6] \\ &= .25\ (.99) + .25\ (.90) + .50\ (.80) \\ &= .88 \quad . \end{aligned}$$

Of course, different weights would yield different proficiency scores, making this a flexible scheme useful for evaluation in quite different settings, although based on the same raw data. Class proficiency scores can be computed for any row or any column of the matrix in Figure 1. With a minor change in the structure, a proficiency score may be derived for a domain defined in terms of any combination of row and column cells in the figure. A mastery criterion score may be specified for the domain or for any cell, class, or combination of cells or classes. The mastery decision, of course, would be based on a comparison of the individual's proficiency score with the mastery standard and a consideration of the importance of relevant types of classification errors. (See Lord, 1977; Millman, 1973; Hambleton & Novick, 1973, for some considerations relevant to mastery classification.)

Estimating task element proficiency scores. The concept outlined above can be used for deriving "true scores" for any subset of a well-defined domain so that any number of uniquely defined task class or domain scores can, in principle, be calculated from data similar to that in the cells of Figure 1. This presents the practical problem of how to obtain the conditional probability values for the task elements. Calculating these values requires knowledge of the relevant item characteristic curves, their respective parameters, and the achievement level θ.

Where estimates of these parameters are available, the task element proficiency scores may be estimated straightforwardly from

$$P_{ig} = P_j\,(\theta_i) \quad , j \in g. \qquad [7]$$

However, practical considerations may require each task element to be defined, not in abstract terms of an infinite number of psychometrically homogeneous items, but rather in operational terms of a finite set of items tried and proven. Suppose just k items can be assembled to measure and define task element t, and the parameters are known. Then task element proficiency is defined by

$$P_{ig} = k^{-1} \sum_{j=1}^{k} P_j(\theta_i) \quad , \qquad [8]$$

which is practically estimated by

$$P_{ig} \cong k^{-1} \sum_{j=1}^{k} \hat{P}_j(\hat{\theta}_i) \quad . \qquad [9]$$

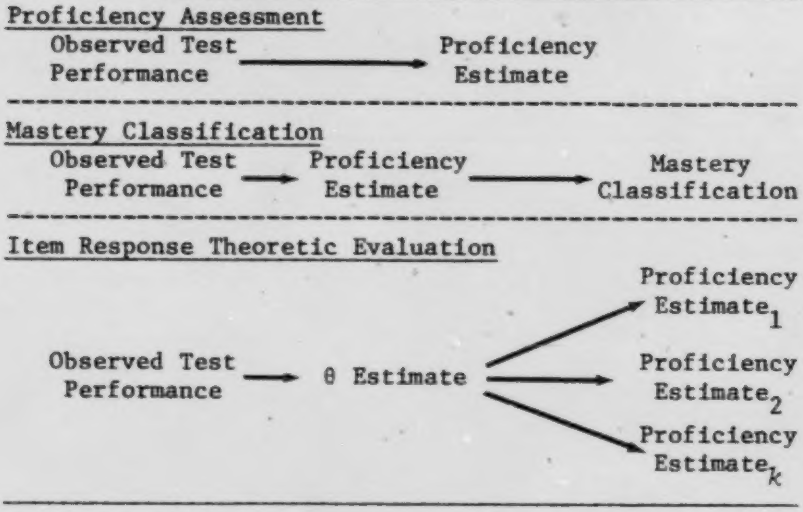
This last formulation is potentially very useful, particularly if it is not feasible to administer any or all of the k items or tasks which define the task element. If the k item response functions are known (or estimated) and accurate information about an individual's achievement level θ is available, task element proficiency can be calculated (or estimated) without observing his/her performance on any of the k critical items. Thus, it may be very useful to measure θ using multiple-choice items and then to use those θ estimates (in conjunction with the response characteristic curves of free-response, production, or performance-type items) to calculate estimates of individual proficiency on task elements which are rather different from the items used as a basis for the estimates.

The distinction between achievement level, θ, and proficiency, P, should be noted; θ is a generalized attribute the value of which is fixed for an individual at a given time. Proficiency, P, is a specific construct which is functionally related to θ, but its value at a given θ level differs as the performance task definition differs. It is customary to estimate P from individual performance on a test considered to reflect a random sampling of items from the specified relevant universe of items; the proficiency estimate thus obtained is not functionally related to proficiency on other related task domains, classes, or elements.

Comparison with other approaches. The performance evaluation paradigm described above can be contrasted with more traditional paradigms. Figure 2 illustrates these differences. At the top of Figure 2 is depicted the very simple paradigm involved in proficiency assessment: A direct proficiency estimate is made on the basis of total score on a criterion-referenced test or performance sample. Implicit in this procedure is one of two assumptions: (1) that all the items or performance tasks are equivalent psychometrically or (2) that they are sampled randomly from the universe of items or tasks which measure the performance proficiency. Either assumption is restrictive for purposes of practical test design.

Figure 2

Traditional and Proposed Paradigms for
Criterion-Referenced Test Interpretation


Proficiency Assessment
Observed Test Performance
Proficiency Estimate
Mastery Classification
Observed Test Performance
Proficiency Estimate
Mastery Classification
Item Response Theoretic Evaluation
Observed Test Performance
θ Estimate
Proficiency Estimate$_1$
Proficiency Estimate$_2$
Proficiency Estimate$_k$

In the center of Figure 2 is illustrated the slightly more complex paradigm underlying traditional procedures for mastery testing. The examinee is classified on the basis of his/her test performance (number correct score), which is perhaps mediated by a proficiency estimate. In Kriewall's formulation, a proficiency estimate (relative true score) would be calculated and compared with the mastery criterion in order to make the mastery decision. Other formulations (e.g., Novick & Lewis, 1974) provide for a mastery decision based directly on the number correct score, with no implicit estimate made of proficiency level. Either formulation involves the same assumptions noted above for the proficiency assessment case.

The use of item response theory for the design and analysis of criterion-referenced tests supports a more general paradigm, such as that shown at the bottom of Figure 2. From item-by-item consideration of the examinee's test or task sample performance, an estimate of θ (his/her functional achievement level on the attribute scale) can be calculated. The θ estimate may then be combined with the known characteristic curve parameters of measures of various critical task elements to yield estimates of proficiency on, not one, but several task elements at once. These estimates are then combined into task class proficiency and/or domain true score estimates. The resulting estimates may be of direct interest, or they may be used as the basis for as many mastery classifications as there are proficiency estimates.

In applications of item response theory, θ is estimated using statistical techniques which consider the parameters of each item's characteristic function and the pattern of the examinee's responses to the items answered. All psychometric information in the examinee's response protocol can be used. Comparable scores can be derived, even though different individuals may take somewhat different tests; there is no need to assume that a test is a random sample from a universe of test items. Differences across items in difficulty and discriminating power can be taken into account in the scoring, and the influence of guessing on test performance can be accounted for and controlled.

A variety of procedures is available for estimating θ from an examinee's responses to a set of test items. These include maximum likelihood estimation (Lord, 1974); Bayesian modal estimation (Samejima, 1969; Urry, 1976); Bayesian sequential estimation (Owen, 1975); and others. Sympson (1976) has provided a review and comparison of some of these methods.

Item response theory also avoids the difficulties encountered when classical test theory is applied to criterion-referenced tests. The classical indices of item difficulty (*p*-value) and item discrimination (item-test correlation) are replaced by the population-invariant threshold and slope parameters of the item characteristic function. The notion of global item difficulty gives way to conditional difficulty. The problem of the irrelevance of norm-referenced test reliability indices to criterion-referenced tests is obviated, since the conditional standard error of estimation (or the test information function) is the fundamental index of measurement precision (Samejima, 1977), the value of which is measured locally as a function of θ.

Given that item response theory is, in principle, broadly applicable to criterion-referenced tests and that θ estimates and item characteristic curves can be quite useful and flexible for domain-referenced proficiency estimation, it remains to be seen what item response models and item parameter estimation procedures are appropriate for use in the context of criterion-referenced measurement.

Item Response Models for Criterion-Referenced Testing

Existing Response Models

The most widely applied item response models are ogives, particularly variants of normal and logistic ogives. Since normal and logistic ogive models can be made nearly indistinguishable by an appropriate scaling transformation, discussion can be limited to consideration of four models based on the logistic ogive. Mathematical descriptions of these are available elsewhere (e.g., Hambleton & Cook, 1977); this discussion will proceed at a conceptual level.

Rasch model. The Rasch model (Wright & Panchapakesan, 1969) is a one-parameter logistic ogive expressing item probability of a correct response as a simple logistic function of the distance between an examinee's location on an attribute scale and the location of the item threshold parameter on the same scale. All item characteristic curves are assumed to have equal slopes. The scale uses real numbers so that θ varies from minus to plus infinity. The actual scaling of θ and of item parameters is theoretically invariant across examinee populations but is *not* invariant over item samples. The scale is in effect determined by the set of items used and by their discriminating power (Urry, 1977). In short, the parameter of an item varies with its context. The Rasch model proper does not admit the effect of guessing on probability correct; and strictly speaking, it is inappropriate for use with multiple-choice items.

A modified Rasch model was proposed by Urry (1970). It has all the features of the basic one-parameter model plus a lower asymptote parameter to accomodate the effects of chance success on probability correct.

Two-parameter logistic model. The two-parameter logistic model (Birnbaum, 1968) is formally similar to the Rasch model, but it specifically incorporates an item characteristic curve slope parameter, which provides for differences in item-discriminating power. It is not technically appropriate for use with multiple-choice items because of the absence of a lower asymptote parameter.

Three-parameter logistic model. The three-parameter logistic model (Birnbaum, 1968) is an extension of the two-parameter model to include a lower asymptote, or guessing, parameter. It is the most general model (Urry, 1977) and can accomodate any response data for which the other models are appropriate, as well as data from responses to multiple-choice items for which the first and second models are inappropriate. The item parameter estimation procedures used in conjunction with it (e.g., Urry, 1976; Wood, Wingersky, & Lord, 1976) in principle are context-free; they are not invariant across examinee populations (Wright, 1977). However, equating transformations are available (Lord, 1975) which mitigate the seriousness of this feature.

A Time-Referenced Response Model

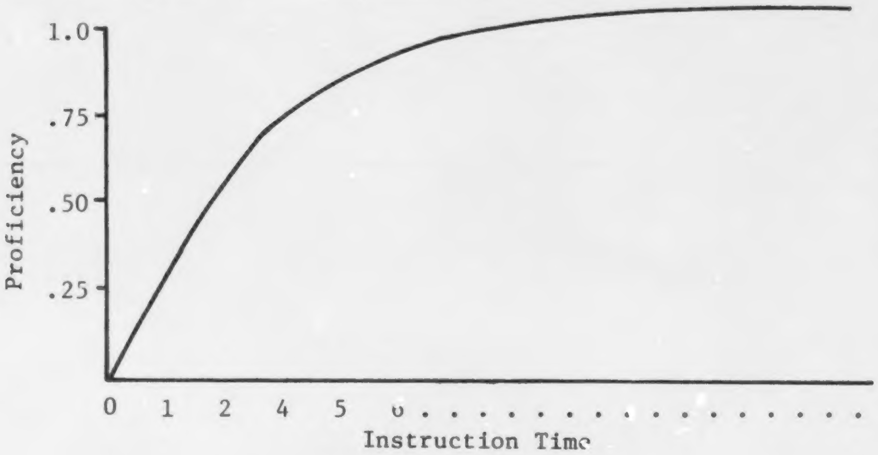
The attribute scales employed in conjunction with the four logistic model variations are all interval scales, with arbitrary origins and units having no inherent objective meaning. They are, in effect, mathematical or statistical abstractions similar to the principal components of a correlation matrix. The scale values themselves have no more direct objective interpretability than do standardized test scores, although they do support indirect objective interpretations (e.g., when used as a basis for estimating proficiency scores as described above). It would seem preferable to employ scales which not only support objective interpretations but which are themselves objectively interpretable.

Time as a metric. It was stated above that achievement level is a function of time or, more specifically, of experience or instruction. It might be useful to exploit this relationship, making time--or some function of it--the scale and unit of measurement basic to item response models used in a criterion-referenced measurement setting. It is well known that achievement in a course of instruction is closely related to the time spent under instruction, be it elapsed time or time-on-task (Bloom, 1974). Carroll (1963) defined aptitude in terms of a time-to-achieve learning criterion, recognizing that individuals differ greatly in the amount of time needed to learn skills or develop competencies.

Bloom (1974) explicitly suggested that time might be useful as a metric for determining human variation in learning settings; his suggestion would make time an outcome or dependent variable. The formulation suggested here employs time as the attribute metric, i.e., as an independent or input variable.

Consider the development of a complex skill being taught in a course of instruction. The group or average "learning curve" for skill proficiency is expressed as a function of time (or trials) and typically has the shape of an asymptotic regression function, such as Figure 3. The function represented by this curve expresses the mean skill proficiency conditional on learning time. If it is known how long a person has been under instruction, this curve can be used to infer his/her proficiency. Over a large number of such inferences, the mean error should be close to zero; but the errors for a given individual may be systemically greater or less than zero due to the effects of individual differences. In order to make accurate proficiency inferences, individual differences must be taken into account; however, the effect of individual differences on proficiency will vary with time.

Figure 3
Skill Proficiency as a Function of Time Under Instruction



If all the relationships involved are understood, proficiency can be predicted as a function of measured aptitude and time under instruction. However, what is of interest is, not *predicting* proficiency, but measuring or estimating an individual's *current* proficiency. For this purpose the latent attribute notion is useful. Using the rationale and methods of latent trait test theory: (1) proficiency can be modeled as a function of time; (2) an individual's location on the latent attribute scale, or "time-equivalent" level, can be measured; and (3) the resulting models and measures can be used to estimate task element, class, and domain proficiency from performance on a criterion-referenced test.

Proficiency modeling. Proficiency modeling methods can closely parallel the response modeling procedures of item characteristic curve theory. For a given item or performance task, it can be assumed that proficiency (probability of a correct response or successful performance) is a monotonic increasing function of time. Some mathematical function can be found which closely fits sampled empirical data for the regression of proficiency on time. This function may be an ogive; an asymptotic regression function, such as that depicted in Figure 3; or perhaps a polynomial of conveniently low degree.

The functional form may be chosen a priori, or the choice of form may be based on available data. Task elements and test items may be characterized by parameters of the function chosen. These parameters are analogues of the parameters of item characteristic curves, except that they have reference to a time metric. Each task or item may have location, scale, and lower asymptote parameters. The location parameter is a threshold on the time scale, the scale parameter refers to the rate of proficiency increase as a function of location, and the lower asymptote parameter may vary above zero to reflect pre-instruction performance levels or the effect of random guessing on response probabilities.

Estimating person location. Each person may be characterized by a location parameter in the same metric as the item threshold--time. This parameter is conceived here as a "time-equivalent" value, which differs from an individual's actual time under instruction as a result of individual differences. The probability of a correct item response or successful task performance is viewed as a function of the difference between the person's time-scale location and the item's threshold parameter. The outcome of a criterion-referenced test thus may be an estimate of the person's time-equivalent scale location, which may be used in turn as the basis for proficiency inferences.

The curve in Figure 3 may serve as one form for modeling proficiency. It has the mathematical form:

$$Y = 1 - (1-c)\, e^{-ad} \quad , \qquad [10]$$

where $Y = P(\theta)$ = proficiency;
$a > 0$ is the item rate parameter;
b is the item threshold parameter;
c is the lower asymptote parameter $(0 < c < 1)$; and
$d = (\theta - b)$ is a difference measure $(d > 0)$.

Practical item parameter estimation procedures for this model have not been explicitly developed. However, they can be based on empirical data relating proportion of correct or successful responses to a direct measure of time (e.g., instruction time, elapsed time, time on task). Similar procedures can be devised for using time and time-equivalent metrics in conjunction with the logistic response models referred to earlier.

Estimating proficiency from performance. Whether a time-referenced or a more usual latent attribute metric is used, once an item or task has been calibrated, it can become a highly useful component of a bank of items and tasks which serve to measure and/or to define task elements, classes, and domains. The parameters of each of these are indices of the psychometric properties of the item or task and are potentially as useful in criterion-referenced testing as they have been in applications of latent trait theory to norm-referenced tests by virtue of their invariance properties. They will be useful in (1) the design of criterion-referenced tests for special purposes, (2) the analysis of such tests (particularly those used in large-scale testing programs), and especially (3) the objective interpretation of individual examinee performance on a criterion-referenced test.

The items and tasks in the bank can all be calibrated with respect to the attribute metric. If such calibration has been accomplished, the bank can contain a mixture of both test items and performance tasks. Estimates of θ based on a person's responses to a subset of the bank can be used to estimate proficiency on any other subset. This means that once the item/task bank is constructed, a person's proficiency at specified calibrated performance tasks can, in principle, be estimated accurately from his/her performance on a test composed of traditional type test items, including multiple-choice test items. This is potentially of great value in circumstances in which time and cost constraints do not permit evaluating performance by administering samples of actual performance tasks.

Conclusion

Outlined above is a conceptual scheme for using the methods of latent trait theory to estimate an examinee's location on a continuous scale of achievement and to infer domain true score from the location estimate. At least two difficulties are immediately apparent, even after the problem of fitting response models to criterion-referenced test items has been solved. One problem is that of bias of estimation; the other is that of dimensionality of the achievement attribute.

Bias. The accuracy with which achievement level (θ) is estimated necessarily will affect the accuracy of the inference from θ to proficiency. To the extent that the estimator of θ is biased, proficiency estimates based on it will have systematic error components. As Lord (1976) has observed, in norm-referenced testing such bias is not a serious problem, particularly if the bias is linear, since it does not affect the ordering of the examinees. In criterion-referenced applications, however, the estimate of θ is to be interpreted objectively by making inference from it to what the examinee can do. The elimination of bias is therefore critical in this context.

Bias may result from (1) use of an inappropriate form for the item response model, such as using the Rasch model in conjunction with multiple-choice items; (2) inaccurate estimation of the parameters of item response characteristic curves; (3) the use of biased estimation techniques; and (4) poor test design, e.g., poor choice of item difficulty. Minimization of bias thus requires careful attention to, among other things, these four sources of bias. The first three can be minimized by (1) using appropriate models, (2) using accurate item parameter estimation methods and sufficient item calibration data, and (3) avoiding the use of estimation techniques known to be biased. The biasing influence of poor test design can be nearly eliminated by the use of adaptive testing procedures to tailor item difficulty to the individual examinee.

Dimensionality. As presently developed, the item response models and methods referred to above are applicable only when the latent trait underlying examinees' item responses is unidimensional. For a content domain characterized by a linear hierarchy of skills, the items which measure those skills, in all likelihood, have sufficient response homogeneity (across task elements and classes) to permit the application of currently available methods. Where the hierarchy is non-linear, however, blind use of these methods may be hazardous. In such a case, it may be possible to partition the domain into linear hierarchies and to use the latent trait estimation methods separately in each one. However, a more satisfactory approach will be to account for each branch of the hierarchy by the use of multidimensional response models such as those proposed by Sympson (1977).

References

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

Bloom, B. S. Time and learning. American Psychologist, 1974, 29, 682-688.

Carroll, J. B. A model of school learning. Teachers' College Record, 1963, 64, 723-733.

Glaser, R., & Nitko, A. J. Measurement in learning and instruction. In R. L. Thorndike (Ed.), Educational Measurement. Washington, DC: American Council on Education, 1971.

Hambleton, R. K., & Cook, L. L. Introduction to latent trait models and their use in analysis of educational test data. Journal of Educational Measurement, 1977, 14, 75-96.

Hambleton, R. K., & Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

Harris, C. W. Some technical characteristics of mastery tests. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), Problems in criterion-referenced measurement. Los Angeles, CA: UCLA Center for the Study of Evaluation, 1974.

Hively, W., Patterson, H. L., & Page, S. A "universe-defined" system of arithmetic tests. Journal of Educational Measurement, 1968, 5, 275-290.

Kriewall, T. E. Aspects and applications of criterion-referenced tests. Paper presented at the 1972 Convention of the American Educational Research Association, 1972. (ERIC document ED 063 333)

Lord, F. M. Estimation of latent ability and item parameters when there are omitted responses. Psychometrika, 1974, 39, 247-264.

Lord, F. M. A survey of equating methods based on item characteristic curve theory (RB-75-13). Princeton, NJ: Educational Testing Service, 1975.

Lord, F. M. Discussion. In W. A. Gorham, J. F. Gugel, F. M. Lord, C. Jensema, F. L. Schmidt, & V. W. Urry, Computers and testing: Steps toward the inevitable conquest (PS-76-1). Washington, DC: U.S. Civil Service Commission, Personnel Research and Development Center, 1976.

Lord, F. M. Practical applicatons of item characteristic curve theory. Journal of Educational Measurement, 1977, 14, 117-138.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

Millman, J. Passing scores and test lengths for domain referenced measures. Review of Educational Research, 1973, 43, 205-216.

Novick, M. R., & Lewis, C. Prescribing test length for criterion-referenced measurement. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), Problems in criterion-referenced measurement. Los Angeles, CA: UCLA Center for the Study of Evaluation, 1974.

Owen, R. J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of the American Statistical Association, 1975, 70, 351-356.

Samejima, F. Estimation of latent ability using a response pattern of graded scores. Psychometric Monograph, No. 17, 1969.

Samejima, F. A use of the information function in tailored testing. Applied Psychological Measurement, 1977,1, 233-247.

Sympson, J. B. Estimation of latent trait status using adaptive testing procedures. Proceedings of the 18th Annual Conference of the Military Testing Association. Pensacola, FL: Naval Education and Training Program Development Center, 1976, 466-487.

Sympson, J. B. A model for testing with multi-dimensional items. Paper presented at Computerized Adaptive Testing '77 Conference. Minneapolis, July 1977.

Urry, V. W. A monte-carlo investigation of logistic mental test models. Unpublished doctoral dissertation, Purdue University, 1970.

Urry, V. W. Ancillary estimators for item parameters of mental test models. In W. A. Gorham, J. F. Gugel, F. M. Lord, C. Jensema, F. L. Schmidt, & V. W. Urry, Computers and testing: Steps toward the inevitable conquest (PS-76-1). Washington, DC: U.S. Civil Service Commission, Personnel Research and Development Center, 1976.

Urry, V. W. Tailored testing: A successful application of latent trait theory. Journal of Educational Measurement, 1977, 14, 181-196.

Wood, R. L. Wingersky, M. & Lord, F. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters (RM 76-6). Princeton, NJ: Educational Testing Service, 1976.

Wright, B. D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.

Wright, B. D., & Panchapakesan, N. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969, 29, 23-48.

UNIFACTOR LATENT TRAIT MODELS APPLIED TO MULTIFACTOR TESTS: RESULTS AND IMPLICATIONS

MARK D. RECKASE
UNIVERSITY OF MISSOURI--COLUMBIA

An assumption basic to most latent trait models is that the hypothetical variable measured by a test can be described in a one-dimensional latent space (Lord & Novick, 1968; Whitely & Dawis, 1974). This assumption can be approximately met for many psychological tests constructed using traditional item and factor analysis procedures. However, if latent trait theory models are to be used in the construction of achievement tests, the effect of violating the unidimensionality assumption must be determined.

Achievement tests are not usually constructed using methodology designed to yield factor pure measures. Instead, a table of specifications is constructed indicating the relative content emphasis and level of behavioral objectives to be measured by the instrument. Items are then written to match the specifications. The tests produced in this way seldom measure a single trait and often will be factorially complex.

The effect of violating the unidimensionality assumption of latent trait models has not been extensively researched. Only four studies have been found in the literature relating the factor structure of tests to latent trait models (Forbes & Ingebo, 1975; Hambleton, 1969; Reckase, 1972; Ryan & Hamm, 1976), although numerous authors have discussed the problem (Hambleton, Swaminathan, Cook, Eignor, & Gifford, 1977; Lord & Novick, 1968; Whitely & Dawis, 1974). Yet the importance of the assumption is indicated by the number of researchers who use it as a reason for lack of fit to the models (e.g., Kifer & Bramble, 1974; Rentz & Bashaw, 1971).

Of the four studies found in the literature, only two (Hambleton, 1969; Reckase, 1972) directly manipulated the factor structure of simulation data to discover the effects on the models; but all four applied latent trait models to actual multidimensional test data. In Hambleton's (1969) study, simulated tests were constructed containing 1 or 5 items from one factor, while the rest of the items in 15- or 30-item tests were from a second factor. The inserted items were found to cause lack of fit of the simple logistic model to the tests; the greater the proportion of odd items, the worse the fit. Reckase (1972) constructed one-, two-, and three-factor simulated 30-item tests with an equal number of items on each factor. When the simple logistic model was checked for fit, the one-factor test fit well; the three-factor test fit moderately well; and the two-factor test did not fit at all. The moderate fit of the three-factor test suggests that the simple logistic model may be somewhat tolerant to violations of the unidimensionality assumption when a number of factors are present.

In the real data studies, both Hambleton (1969) and Reckase (1972) found an increasingly poor fit of the latent trait models as the number of factors increased; however, the results were clouded by the presence of violations of zero guessing and equal discrimination assumptions. In a different kind of study, Forbes and Ingebo (1975) subjectively sorted a seventh-grade mathematics test into three homogeneous item sets. The total test and the subtests were then calibrated separately using the simple logistic model. The results showed that the ranking of the items on difficulty did not change, regardless of the set of items with which they were calibrated; and the parameter estimates for the subsets differed by an approximately constant value from those obtained using the full tests. On the basis of these results, the authors concluded that sorting the items into homogeneous sets was not necessary before the tests were calibrated.

Ryan and Hamm (1976) attacked the problem of unidimensionality from another direction. They analyzed eight tests using both the simple logistic model and principal components factor analysis. On the basis of these analyses, items were selected for each test that either fit the logistic model or loaded on the first factor. The two sets of selected items were again factor analyzed, and the proportion of variance in the first factor was compared. The items selected by the simple logistic model had a first factor accounting for slightly more variance than the original test, while items selected by the factor analysis substantially increased the size of the first factor. These results indicate that the simple logistic model does not select items on the basis of factor purity. Rather, it uses a more complex criterion, which is probably related to guessing and discrimination effects as well as to factor structure.

The results of these four studies indicate that the factor structure of a test will affect the fit of latent trait models in a detrimental way. However, little information is available concerning the factorial complexity of achievement tests; and even less is known about the interaction between univariate latent trait models and factorially complex tests. This leads to a series of interesting questions: (1) Will one- or three-parameter models be more robust to multidimensionality? (2) What trait is measured by the ability estimates from latent trait models when they are applied to multifactor tests? (3) Is there any value to be derived from applying latent trait models to tests without a dominant first factor? These and many other questions have not yet been answered. It is hoped that the research reported in this paper will clarify some of these issues, with the ultimate goal of applying the results to the calibration of achievement tests.

Design

Models and Computer Programs

Two latent trait models were selected for use in this research study: the one-parameter logistic (1PL) model and the three-parameter logistic (3PL) model. These models were selected because of their frequent use in research literature and because of their mathematical tractability. Birnbaum (1968) indicated that the logistic item characteristic curve is very similar to the normal ogive item characteristic curve. At the same time, it is much more convenient from a mathematical point of view.

The 1PL model is given by the following expression:

$$P\{x_{ij} = u\} = \frac{e^{(\theta_i - b_j)u}}{1 + e^{(\theta_i - b_j)}}, \qquad [1]$$

where x_{ij} is the response to item j by person i;

u is the score on the item (0,1);
θ is the ability parameter for person i; and
b_j is the difficulty parameter of item j.

The 3PL model is given by:

$$P\{x_{ij} = u\} = c_j + (1 - c_j) \frac{e^{Da_j(\theta_i - b_j)u}}{1 + e^{Da_j(\theta_i - b_j)}}, \qquad [2]$$

where $D = 1.7$;
c_j is the guessing parameter; and

a_j is the discrimination parameter for item j.

The only difference between the models is the lack of a_j, c_j, and D values in Equation 1. The sole purpose of D is to make the logistic model similar to the normal ogive model. It has no other effect. The a_j and c_j parameters are more important. Their absence in Equation 1 indicates that guessing is assumed to be zero and that discrimination is assumed to be 1.0 for all items. Obviously, Equation 2 gives a more realistic model; but this realism is paid for by the need for more complex estimation procedures.

The calibration procedures used in this study were selected after a review of seven 1PL programs and six 3PL programs. The basis for selection was the accuracy of the programs and their usability for the test lengths and sample size required by the study. The 1PL program selected was the unconditional maximum likelihood program developed by Wright and Panchapakesan (1969). The 3PL program selected was the quasi-maximum likelihood procedure developed by Wood, Lord, and Wingersky (1976). A complete description of all the programs is given in Reckase (in prep.).

Data

In order to evaluate the effects of multidimensionality on the 1PL and 3PL models, 10 data sets, varying in the number of factors present, were obtained for analysis. Five of the data sets were generated using a simulation program to match specific factor structures. The other five were obtained from course examinations and the Missouri Statewide Testing Program. Table 1 summarizes the information on the data sets. The tests and simulation data were selected to have varied factor structures, item quality, and sample sizes so that the results would be generalizable.

Analyses

Factor analysis. The first step in analyzing the data sets described in Table 1 was to determine their factor structure. All of the data sets were subjected to three factor analyses: (1) a principal components analysis on phi-coefficients, (2) a principal factor analysis on phi-coefficients, and (3) a principal components analysis on tetrachoric correlations. The analyses were performed to determine the factor structure of the empirical data sets and to confirm the factor structure of the simulation data sets. The analyses of both the phi correlations and the tetrachoric correlations are reported since, in some cases, the tetrachoric correlations did not yield Gramian matrices for the small data sets.

The factor structures obtained for the ST1075, ST1076, and ST0576 tests were then compared to the table of content specifications for these examinations in order to determine if the factors matched the content outline of the test. The number of factors decided upon for these tests was based on a compromise between statistical considerations and subjective content judgements. Factor scores on the first factors were also determined on each data set, based on each of the three factor analytic techniques, and on the varimax-rotated factors, based on the tetrachoric correlations for the two- and nine-factor simulated data sets.

Item parameterization. After the factor analyses were completed, each of the data sets were then analyzed using the 1PL and 3PL programs. The 3PL program was run using the default options to ensure reproducibility of results. The item and ability parameters from these two analyses were saved for further study. Computer time, cost, and the number of poorly fit items were also recorded for each analysis.

Traditional item analysis. The third analysis performed on these data was a traditional item analysis, including descriptive statistics on the raw score distributions and KR-20 reliability estimates. The purposes of this analysis were to determine the quality of the test data in traditional terms and to acquire traditional item statistics as a basis for comparison with the latent trait parameters.

The analyses described thus far served to acquire the data needed to evaluate the two latent trait procedures. Subsequent analyses were performed

Table 1
Description of Data Sets

Test Name	Abbreviation	Sample Size	Description
Missouri Scholastic Aptitude Test/Verbal	MSCATV	3,126	Systematic sample from 58,000 cases from Statewide Testing Program, 1975-76. SCAT Series II Form 2B.
Missouri Scholastic Aptitude Test/Quantitative	MSCATQ	3,126	Systematic sample from 58,000 cases from Statewide Testing Program, 1975-76. SCAT Series II Form 2B.
Measurement course exam on standardized testing	ST1075	208	Final exam administered in October, 1975.
Measurement course exam on standardized testing	ST1076	176	Final exam administered in October, 1976.
Measurement course exam on standardized testing	ST0576	181	Final exam administered in May, 1976.
One-factor rectangular simulation data	150AR	1,000	One factor with loadings of .9 on factor, rectangular distribution of difficulties.
Two-factor simulation data	250AN	1,000	Loading of 0.9 randomly distributed on the two-factor, normal distribution of difficulties.
Nine-factor Spearman simulation data	950ANS	1,000	Dominant first factor with 0.7 loadings. Items randomly distributed to other eight factors with 0.6 loadings, normal distribution of difficulties.
Nine-factor independent simulation data	950AN9	1,000	Items randomly distributed over nine factors, .9 or .0 loadings on factors, normal distribution of difficulties.
Nine-factor independent simulation data	950AN3	1,000	Items randomly distributed over nine factors, .3 or .0 loadings on factors, normal distribution of difficulties.

Note. All tests were 50 items in length.

in order to determine important characteristics of the item parameter estimates, the ability parameter estimates, and the overall fit of the models. The item parameter characteristics were evaluated by comparing the size and relationship of the item parameter estimates, the factor analysis loadings, and traditional item statistics across the different tests.

Comparison of parameter estimates and factor results. The first evaluative analysis performed on the item statistics was a comparison of the parameter estimates and fit statistics from the logistic models with the theoretical and empirical loadings from the factor analyses. Correlational techniques were used for these comparisons. The two-factor simulation data was analyzed first, followed by the two nine-factor simulation data sets, and then the live testing results. The comparisons of the item parameters were supplemented by comparisons between the latent trait ability estimates and the factor scores to give added evidence concerning the factor being measured.

Once the existence of a relationship between the latent trait parameters and the factor loadings had been demonstrated, the variables controlling the size of the relationship were of interest. One variable which was considered to have some influence was the strength of the first factor. To check this hypothesis, the mean and standard deviation of various parameter values were plotted against the size of the first eigenvalue. Regression techniques were then used to determine the mathematical function describing the effect of the eigenvalues.

Factor analysis of item parameter data. Following the analysis comparing the item statistics to the strength of the first factor, a more global analysis was performed. The analysis compared the values obtained from the simple logistic model, the three-parameter logistic model, the traditional item analysis, and the factor analyses to determine general relationships. To do this the item statistics were intercorrelated and the resulting matrix was factor analyzed, using the principal components technique and rotated to the varimax criterion. The distributions of the item statistics were found to be approximately normal and the number of items sufficiently large to justify this analysis.

Fit of the two models. Finally, a measure of fit of the two models used in the study to the item response data was obtained using the following formula:

$$\text{Test Mean Square Fit} = \frac{\sum_{j=1}^{n} \sum_{i=1}^{N} (x_{ij} - P_{ij})^2}{Nn} \qquad [3]$$

where x_{ij} is the score of person i on item j and P_{ij} is the probability of a correct response for that person and item obtained from a latent trait model. This measure was used for descriptive purposes only. It was used rather than matching empirical item characteristic curves to theoretical ones, because the latter comparison is too dependent on the interval size used for deter-

Table 2
Descriptive Data and Results of Factor Analyses by Test

Statistic	Test									
	MSCATV	MSCATQ	ST1076	ST1075	ST0576	150AR	250AN	950ANS	950AN9	950AN3
Mean	29.14	28.67	34.00	35.00	35.00	25.21	25.23	24.84	25.33	25.00
Standard Deviation	9.22	9.40	5.30	4.10	5.00	13.22	12.98	13.50	6.46	3.81
KR-20	0.90	0.90	0.71	0.56	0.66	0.97	0.95	0.96	0.74	0.22
Expected Number of Factors						1	2	9	9	9
Principal Components										
Number of Factors	8	9	20	21	21	4	4	9	9	22
First Eigenvalue	8.90	9.30	4.35	3.05	3.35	21.45	14.70	15.90	4.10	1.55
Principal Factors										
Number of Factors	2	3	9	9	9	3	3	9	9	0
First Eigenvalue	8.15	8.60	3.85	2.55	2.80	20.15	14.3	15.45	3.65	.80
Principal Components (Tetrachoric)										
Number of Factors	8	9	20	22	21	4	2	9	9	22
First Eigenvalue	14.70	15.30	7.70	7.20	5.60	40.70	21.65	24.60	5.65	1.95
Sample Size	3126	3126	176	208	181	1000	1000	1000	1000	1000

Note. The number of factors for all factor analyses is based on the eigenvalue greater than 1.0 rule.

Table 3
Factor Loadings, 1PL Fit,
and 3PL Discrimination for Data Set 250AN

					Tetrachoric Correlations		
	Theoretical Loadings		1PL	3PL	Varimax		Principal Components
Item	Factor 1	Factor 2	Fit	Discrimination	Factor 1	Factor 2	Factor 1
1	9	9	00	02	95	02	76
2	0	9	00	66	-07	90	50
3	0	9	01	133	00	88	54
4	9	0	17	09	89	08	75
5	9	0	02	11	92	03	74
6	0	9	48	186	10	90	64
7	0	9	42	172	03	91	59
8	9	0	22	11	91	02	73
9	9	0	00	10	91	-03	70
10	0	9	57	192	02	91	58
11	0	9	54	176	03	90	58
12	9	0	84	09	90	00	71
13	0	9	29	183	04	91	59
14	9	0	21	08	91	-01	71
15	9	0	52	12	92	02	73
16	0	9	92	165	-01	90	55
17	0	9	83	177	04	91	59
18	9	0	89	13	90	05	74
19	0	9	78	190	03	92	59
20	9	0	97	14	92	01	73
21	9	0	94	15	92	03	74
22	0	9	22	186	03	91	59
23	0	9	35	184	-01	91	55
24	9	0	67	14	90	02	72
25	0	9	25	197	-01	92	56
26	9	0	29	13	91	-01	71
27	9	0	20	12	92	-02	71
28	0	9	92	182	04	91	59
29	9	0	98	14	91	04	75
30	0	9	40	210	01	92	57
31	0	9	61	179	08	90	62
32	9	0	37	13	93	01	74
33	9	0	84	12	92	-01	72
34	0	9	93	210	02	91	58
35	9	0	22	15	91	05	75
36	0	9	70	186	-01	91	55
37	0	9	11	210	05	92	61
38	9	0	12	13	92	-00	73
39	0	9	15	210	00	93	58
40	9	0	12	12	91	-01	71
41	9	0	19	15	91	04	74
42	0	9	37	180	03	89	57
43	0	9	09	210	00	91	56
44	9	0	64	15	90	06	74
45	0	9	12	210	-02	90	54
46	9	0	21	15	92	01	73
47	9	0	20	14	90	03	73
48	0	9	26	193	04	89	58
49	0	9	00	210	-01	89	54
50	9	0	00	14	95	04	77

Note. All values presented without decimal points.

mining the empirical curves. The mean square fit was computed for each data set using the two latent trait models. The mean square fit was also compared to the percent of variance accounted for by the first factor of the tetrachoric factor analysis to determine if fit was related to factorial complexity.

Results

So that the results of the study will be presented in a meaningful fashion, they will be reported separately for several of the data sets--starting with one of the simpler data sets, the two-factor simulation data. The analyses on the more complex live-testing data will subsequently be presented, followed by analyses that cut across all of the data sets used.

First, however, general summary statistics for the 10 data sets are presented in Table 2. These include the mean, standard deviation, KR-20 reliability, theoretical number of factors, obtained number of factors from each factor analysis, size of first eigenvalue from each factor analysis, and the sample size. Note that the number of factors determined by the tetrachoric correlation analysis yielded the most accurate results for the simulation data. Also the data sets varied widely in factorial complexity, traditional reliability, and sample size.

The Latent Trait Dimension

The major question to be addressed in the initial analyses was, What is the dimension that is being scaled by the latent trait models? In other words, on what dimension does discrimination between the abilities of various persons take place? To give an initial answer to this question, the simplest data set containing more than one dimension was used, 250AN (see Table 1).

The factor loadings used to generate this data set along with the 1PL fit values, the 3PL discrimination estimates, and factor loadings from rotated and unrotated tetrachoric correlation factor analyses are presented in Table 3. This table shows that the 3PL discrimination parameter estimates corresponded very closely to the second factor of both the theoretical and empirical loadings. A negative relation was observed between the discrimination estimates and the first principal component. The correlations describing these relationships are presented in Table 4.

Table 4
Correlations between Factor Loadings, 1PL Fit, and 3PL Discrimination for 250AN

Variable	2	3	4	5	6	7
1. Theoretical Loading Factor 1	-1.00	-.08	-.97	1.00	-1.00	0.96
2. Theoretical Loading Factor 2		.08	.97	-1.00	1.00	-0.96
3. 1PL Fit			.11	-.07	.09	-0.01
4. 3PL Discrimination				-.96	.97	-.90
5. Varimax Factor 1					-1.00	.97
6. Varimax Factor 2						-.96
7. Principal Component						

The correlations between these values present some other interesting information. First, the 1PL chi-square probability of fit value did not correlate significantly with any of the other statistics. This indicates that neither variation in discrimination nor the multifactor nature of the data can be used to explain lack of fit for this data set. Second, the varimax-rotated loadings and the principal component loadings both correlated highly with the theoretical loadings, indicating the accuracy of the analysis.

Table 5
Correlations between Ability Estimates,
Raw Scores, and Factor Scores for the 10 Data Sets

Data Set	Ability Estimate	Variable Raw Score	 θ	Phi Principal Component	Tet Principal Component	Tet 1	Varimax 2
250AN	3PL	.59		.59	.56	.29	.56
	1PL	.98	.66	.98	.97	.71	.71
950	3PL	.62		.82	.67	.74	
	1PL	.99	.62	.72	.72	.44	
950AN9	3PL	.93		.93	.94		
	1PL	.98	.96	.98	.98		
150AR	3PL	.97		.97	.98		
	1PL	.95	.99	.95	.97		
950AN3	3PL	.71		.36	.41		
	1PL	1.00	.71	.25	.33		
MSVAT1	3PL	.98		.99	.99		
	1PL	.99	.97	.98	.98		
MSCAT2	3PL	.97		.98	.98		
	1PL	.99	.96	.97	.97		
ST1075	3PL	.83		.89	.32		
	1PL	.99	.85	.89	.29		
ST0576	3PL	.88		.91	.87		
	1PL	.99	.90	.93	.88		
ST1076	3PL	.89		.94	.91		
	1PL	.98	.90	.88	.86		

The relation between the discrimination estimates and the second factor would seem to indicate that the 3PL model was measuring the second factor on the test. This hypothesis can be checked by correlating the ability estimates from the 3PL procedure and the factor scores from the factor analytic procedures. This information is presented in Table 5. The results presented here were somewhat surprising. Although the 3PL ability estimates were clearly more closely related to the second rotated factor than the first, the correlation with the second factor was surprisingly low (.56). It was about the same size as the correlation with the first principal component and the raw score. The 1PL

Table 6

Factor Loadings, 1PL Fit, and 3 PL Discrimination for 950AN

Item	Theoretical Factor Loadings I	II	III	IV	V	VI	VII	VIII	IX	1PL Fit	3PL Discrimination	Tetrachoric Varimax Factor 1	Phi Principal Component Factor 1
1	0	0	0	0	0	0	0	9	0	52	05	-00	07
2	0	0	9	0	0	0	0	9	0	92	193	-00	15
3	0	0	0	0	9	0	0	0	0	72	16	-01	26
4	0	0	0	0	0	0	9	0	0	21	01	00	-02
5	0	0	0	9	0	0	0	0	0	71	04	-04	-01
6	0	0	9	0	0	0	0	0	0	70	193	01	21
7	0	0	9	0	0	0	0	0	0	71	193	-02	18
8	0	0	0	0	0	0	0	0	9	99	193	91	62
9	0	0	0	0	0	0	0	9	0	40	03	00	07
10	0	0	0	0	0	0	0	9	0	88	03	04	07
11	0	0	0	0	0	0	0	9	0	02	03	04	07
12	0	0	0	0	0	0	0	0	9	33	193	93	69
13	0	0	0	0	0	9	0	0	0	20	03	-05	-01
14	9	0	0	0	0	0	0	0	0	64	06	-03	08
15	0	0	0	0	0	0	9	0	0	73	01	-01	-00
16	0	9	0	0	0	0	0	0	0	64	13	-01	11
17	0	0	0	0	9	0	0	0	0	42	24	05	35
18	0	0	0	9	0	0	0	0	0	08	05	-03	01
19	0	0	0	9	0	0	0	0	0	28	04	-02	-01
20	0	0	0	0	0	0	0	0	9	29	193	92	70
21	0	0	0	0	0	9	0	0	0	25	06	02	04
22	0	0	0	0	9	0	0	0	0	24	24	06	35
23	9	0	0	0	0	0	0	0	0	99	06	-02	10
24	0	0	0	0	0	0	0	0	9	85	193	94	70
25	0	9	0	0	0	0	0	0	0	82	13	00	11
26	0	9	0	0	0	0	0	0	0	55	12	-01	12
27	0	0	0	0	0	0	0	0	9	43	193	93	69
28	0	0	9	0	0	0	0	0	0	08	15	01	21
29	0	0	0	0	0	9	0	0	0	40	04	-01	02
30	0	9	0	0	0	0	0	0	0	63	13	05	14
31	0	0	0	0	0	0	9	0	0	56	01	00	00
32	9	0	0	0	0	0	0	0	0	39	08	02	11
33	0	0	0	0	0	0	0	0	0	96	13	-02	21
34	0	0	0	0	9	0	0	0	0	61	26	03	33
35	0	0	0	0	0	0	9	0	0	04	01	-00	00
36	0	0	0	0	0	0	0	9	0	19	02	00	05
37	0	0	0	9	0	0	0	0	0	24	07	04	03
38	0	0	9	0	0	0	0	0	0	31	15	06	22
39	9	0	0	0	0	0	0	0	0	86	09	04	14
40	0	9	0	0	0	0	0	0	0	62	15	01	12
41	0	0	0	0	9	0	0	0	0	18	23	02	32
42	0	0	0	9	0	0	0	0	0	58	06	04	04
43	0	0	0	0	0	0	9	0	0	37	02	02	03
44	0	0	0	0	0	0	0	0	9	77	193	93	61
45	0	0	0	0	0	9	0	0	0	19	05	-00	02
46	0	0	0	0	0	0	9	0	0	71	04	07	05
47	0	0	0	0	9	0	0	0	0	60	22	01	27
48	0	0	0	0	0	9	0	0	0	88	06	07	05
49	9	0	0	0	0	0	0	0	0	74	10	02	11
50	9	0	0	0	0	0	0	0	0	81	07	-00	06

Note. All values presented without decimal points ability estimates, on the other hand, correlated highly with the raw score--the raw scores being sufficient statistics for the ability estimates--and the first factor scores, and equally well with the two sets of rotated factor scores. The results are exactly what would be expected if the 1PL estimates were based on the sum of the scores on the two factors. On the basis of these results, it would seem that the 3PL model was estimating the second factor (though rather poorly), while the 1PL model was estimating the sum of the two factors.

A more stringent test of this hypothesis was obtained from the analysis of the 950AN9 data set. Table 6 shows the generating factor structure for this test, the 1PL fit, the 3PL discrimination values, and the first principal component (phi) and varimax factors (tet) for the data set. Although not as clear as the results for the previous data, these results showed that the discrimination values were strongly related to the ninth theoretical factor. Only items 2, 6, and 7 distorted the relationship that is present. A similar relationship existed with the empirically derived factor loadings. The correlations describing these relationships are shown in Table 7.

Table 7
Correlations between Factor Loadings, 1PL Fit, and 3PL Discrimination for Data Set 950AN9

Variable	2	3	4	5	6	7	8	9	10	11	12	13
1. Theoretical Factor 1	-12	-14	-12	-14	-12	-14	-12	-14	29	-18	-14	-14
2. Theoretical Factor 2		-12	-11	-12	-11	-12	-11	-12	15	-14	-13	-09
3. Theoretical Factor 3			-12	-14	-12	-14	-12	-14	12	32	-14	04
4. Theoretical Factor 4				-12	-11	-12	-11	-12	-18	-17	-14	-27
5. Theoretical Factor 5					-12	-14	-12	-14	-08	-10	-12	24
6. Theoretical Factor 6						-12	-11	-12	-17	-17	-13	-25
7. Theoretical Factor 7							-12	-14	-12	-21	-13	-30
8. Theoretical Factor 8								-12	-15	-18	-12	-18
9. Theoretical Factor 9									11	78	100	88
10. 1PL Fit										25	11	14
11. 3PL Discrimination											78	78
12. Varimax Factor 1												89
13. Principal Component 1												

Note. Significant values are underlined. Decimal points omitted.

The correlational data reinforce the results of the 250AN analysis. The fit of the 1PL model showed little correlation with any of the values (there was one barely significant correlation with Theoretical Factor 1) and the 3PL discrimination correlations reflected the subjective judgements made on the basis of Table 6. It should be noted that the first varimax factor was obviously Theoretical Factor 9 and that the first principal component was also highly related to this factor.

The correlations with the factor scores yielded a less interpretable result than for the 250AN data (see Table 5). As expected, the 1PL ability estimates correlated highly with the raw scores and had moderate correlations with the principal components and varimax scores. This result was probably due to the fact that these three factor analytic solutions estimated Theoretical Factor 9, while the 1PL estimates were related to sums of all the factors. The 3PL estimates were related to the empirical estimates of Factor 9; though again the correlations were lower than expected, but higher than for the 250AN data set.

Although these data sets are informative for the purposes of understanding the models, they are hardly realistic. Few tests are made up of perfectly independent subsets. A more realistic situation is to have a common first factor--probably verbal ability--with a number of specific factors. This type of test is simulated using data set 950ANS. The correlations with the theoretical loadings are no longer meaningful for this data set, since all the items have .7 loadings on the first factor, yielding zero variance. The correlations with the factor scores still indicate the factor being measured by the models, however (see Table 5). In this case, both models clearly estimate the first factor of the test.

Table 5 also presents the correlations with the raw scores and factor scores for the other seven data sets used in the study. In all cases except two, the results show that the two models were measuring the first factor of the test. This is true even in the case of the three classroom exams with first factors accounting for less than 10% of the total test variance. The exceptions were data sets 950AN3 and ST1075. In the former case, the test was so poor (KR-20 = .22) that neither procedure estimated anything. In the latter case, the tetrachoric correlations yielded unstable results for the small sample size, resulting in the low correlations.

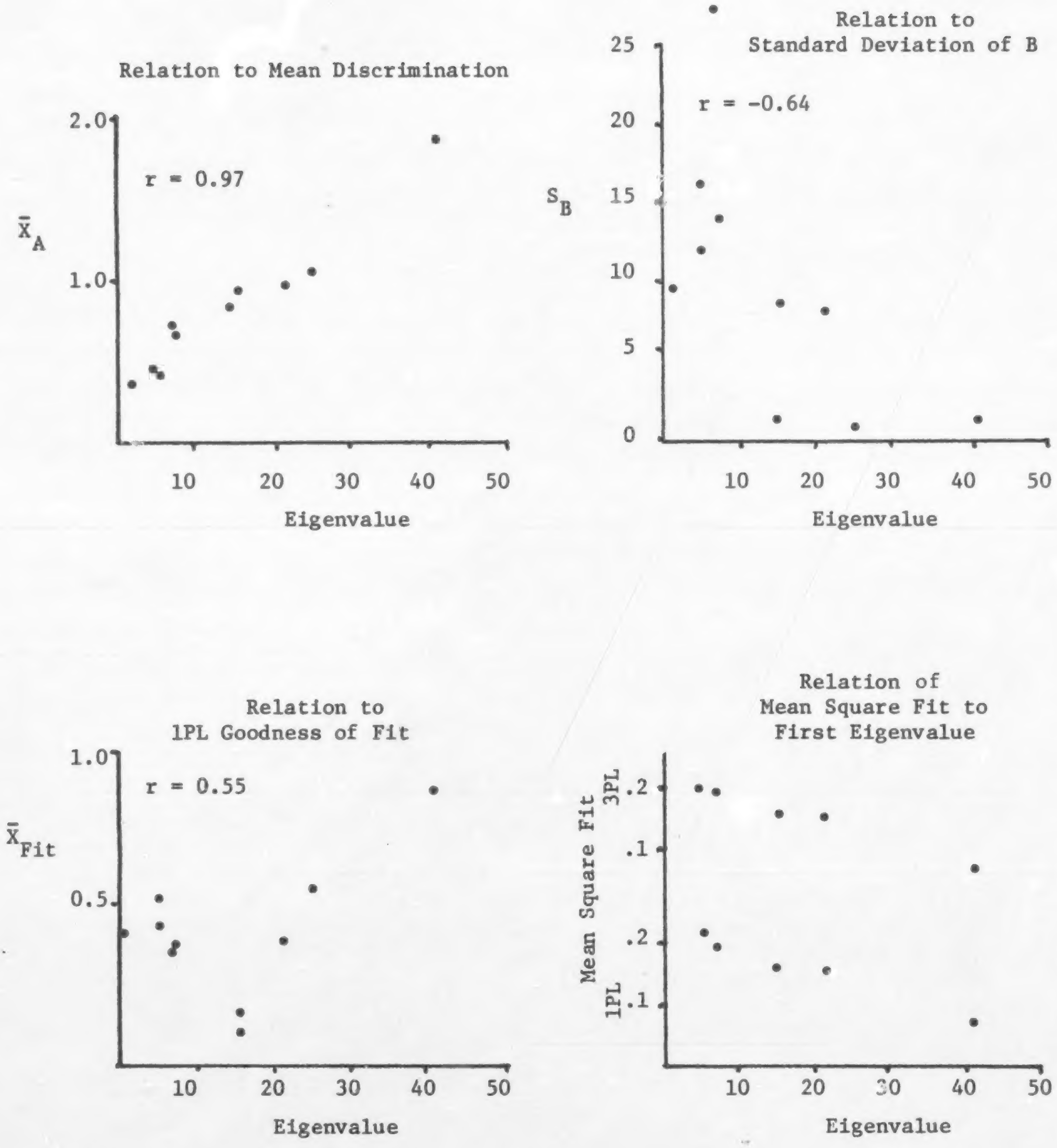
In summary, these results show that the 3PL model estimates one factor when independent factors are present, while the 1PL model estimates the sum of the factors. If a dominant first factor is present, it is estimated by both models.

Adequacy of Trait Estimation

Thus far, the results show what is being measured by the two models; how well the trait is being measured will be evaluated relative to the size of the first factor of each data set. To do this the mean 3PL discrimination, the standard deviation of the 3PL difficulty, the mean probability of fit for the 1PL model, and the mean square fit for the two models were plotted against the first eigenvalues from the tetrachoric factor analyses. These plots are shown in Figure 1.

The relationship shown in the figure for the mean 3PL discrimination and the eigenvalue is very strong. As the first eigenvalue increased, the

Figure 1
Relationship of Eigenvalues
to Selected Item Statistics



average discrimination also increased. This is the expected result because Lord and Novick (1968, chap. 16) and Lord (1952) have shown that the discrimination parameter estimate can be computed from the first factor loadings when the data are unidimensional. The results shown here merely serve to generalize the findings to multivariate data.

The plot of the standard deviations of the 3PL difficulty estimates showed a more interesting relationship. As the size of the first eigenvalue decreased, the standard deviation of the difficulty estimates increased. This result indicates less stability in the 3PL estimation procedure for multifactor tests, since the range of the traditional difficulty indices for the data sets is approximately the same. When the parameters were poorly estimated, the difficulty values tended to become extreme, inflating the standard deviation. The linear correlation in this case was -.64, although the relationship seemed to be hyperbolic.

The 1PL goodness of fit plot yielded a more complex result than the previous two. One set of points fell approximately along a line from the lower left to the upper right. These points come from the data sets with relatively dominant first factors. The other set of points clustered at the left of the graph comes from the data sets with weak first factors. Thus when a dominant first factor is present, fit is directly related to the size of the first eigenvalue. When a dominant factor is not present, the size of the eigenvalue does not seem to be related to fit of the 1PL model.

The relationship between the mean-square fit statistic and the size of the first eigenvalue was much clearer. As the size of the first eigenvalue increased, the deviations from fit decreased in almost a linear fashion from the value of .25 for a 0 eigenvalue to 0 for an eigenvalue of 50. The end points correspond to items with no discrimination in the first instance and perfect discrimination in the second.

Factor Analysis of Item Statistics

The final analysis performed on the data was the correlation of the item statistics from the classroom final exams, followed by a principal components factor analysis of the resulting correlation matrix. Before this analysis was performed, the distribution of the 150 values of each of the item statistics was formed to check for normality. Since the distributions were all unimodal and approximately symmetrical, it was felt that performing this analysis was reasonable. The item statistics from the simulated data sets often gave distributions that were very restricted in range and skewed, making them less appropriate for these techniques. The factor analysis of the intercorrelations yielded three factors; the loadings are presented in Table 8.

The first factor had high loadings on the phi factor loadings and traditional point-biserial discrimination, and moderate loadings on the tetrachoric loadings and 3PL discrimination. The factor is obviously a discrimination factor. The lower loading for tetrachoric loadings is due

Table 8
Varimax Factor Loadings from Intercorrelation
of Item Statistics for the ST Series Tests

Statistic	Factor 1	2	3
Principal Component 1	0.98	0.07	0.02
Principal Factor 1	0.97	0.07	0.03
Tet Principal Factor 1	0.69	-0.01	0.17
3PL Discrimination	0.51	-0.45	-0.14
3PL Difficulty	0.27	-0.46	0.15
3PL Guessing	0.22	0.40	-0.68
1PL Easiness	0.12	0.96	0.03
1PL Fit	0.26	0.22	0.77
Proportion Correct	0.12	0.94	0.02
Point Biserial	0.94	-0.06	-0.04
Eigenvalue	3.73	2.44	1.14

Note. This analysis was performed on 150 items from the three ST1075, ST0576, ST1076 tests.

to the instability of the tetrachoric correlations with the relatively small sample sizes. The variance of the 3PL discrimination was split between Factors 1 and 2. Factor 2 had high loadings on the simple logistic log easiness parameter and proportion correct, with moderate loadings on A and B, the 3PL discrimination and difficulty factors. This factor has been labeled a difficulty factor. The fact that 3PL discrimination loads on both the discrimination and difficulty factors shows the same type of dependence between these parameters as that discussed by Lord (1975). The moderate loading for the difficulty statistic is also due to instabilities caused by the small sample sizes and the weak first factors for these tests. The third factor in this analysis had two high loadings on the 3PL guessing parameter and 1PL fit index. This indicates that the goodness of fit to the simple logistic model is more closely related to variations in guessing than to variations in discrimination.

One conclusion that can be drawn from this analysis is that the 3PL discrimination parameter is related to the size of the loadings of the first factor of the tests, as well as to the distribution of item difficulties. Also, the fit of the 1PL model to a test is more closely related to the guessing level of the items than the loadings of the first factor.

Discussion and Conclusion

The purpose of this study has been to evaluate the one- and three-parameter logistic models for use with multivariate data. To accomplish this evaluation five simulated and five real data sets were selected and analyzed using the 1PL and 3PL techniques, factor analysis, and traditional item analysis techniques. The results of these techniques were compared to answer three questions. First, what component of the tests is being measured by the two models? Second, does the size of the first factor control the estimation of the parameters of the two models? And, third, what is the relationship between the factor analysis, latent trait, and item analysis parameters?

To answer the first question, the 3PL discrimination parameter estimates and 1PL probability of fit statistics were compared to theoretical and empirical factor loadings; and the 1PL and 3PL ability estimates were compared to the factor scores. The results of these analyses fall under two cases. Case 1 includes the analyses of the data sets containing more than one independent factor; the 3PL procedure seemed to select one factor and discriminate among ability levels on this factor, while ignoring the other factors. The correlations between the ability estimates reflected this, but were lower than expected (data sets 250AN and 950AN), possibly because the 3PL estimates were less stable than the factor analytic results. The 1PL procedure seemed to estimate the sum of the factors, which is consistent with the fact that the total score (the sum of the subscores) is a sufficient statistic for the ability estimates. Case 2 includes data sets which have first factors that are large, relative to the other factors in the test. In this case, both methods measured the first factor; the correlations with the factor scores were almost identical, leaving little basis for choice between the two techniques except cost. On a cost per analysis basis, the 1PL model is clearly preferred.

The answer to the second question, Does the size of the first factor control the estimation of the parameters? is clearly yes. The size of the 3PL discrimination parameter estimates, the stability of the 3PL difficulty estimates, the 1PL probability of fit, and the mean squared deviations for both the models showed strong relationships with the first eigenvalues. If a minimum 3PL discrimination parameter of .60 is desired, these relationships indicate that the first factor should have a phi-coefficient-based eigenvalue of 10 or greater for these 50-item tests, or account for at least 20% of the total variance. Tests with smaller first factors will still give reasonable ability estimates, as indicated by the correlations with the factor scores; but the item parameters will be unstable.

The relationships between the item statistics from the factor analyses, latent trait analyses and traditional item analyses were determined by a factor analysis of the correlations between the statistics for estimates from 150 items. The analysis indicates that three factors are present in these parameters: discrimination, difficulty, and guessing. The 3PL and 1PL difficulty estimates and the traditional difficulty loaded on the difficulty factor, as would be expected; but the 3PL discrimination also loaded on this factor, replicating a finding by Lord (1975) showing that the difficulty and discrimination parameters are dependent. The 3PL discrimination parameter also loaded with traditional discrimination and the factor loadings as it should. The 1PL fit statistic loaded with 3PL guessing, suggesting that guessing is the major component in lack of fit, not variations in discrimination or multidimensionality. This hypothesis is supported by the lack of correlations between the fit statistic and the factor loadings in the earlier analyses.

Thus, these results show that the 1PL and 3PL models estimate different abilities when a test measures independent factors; but that both estimate the first principal component when it is large relative to the other models, even when the first factor accounts for less than 16% of the test variance. Item calibration results, however, will tend to be unstable. For acceptable calibration, the first factor should account for at least 20% of the test variance. If a factor other than the first factor is of interest, factor pure subtests should be formed and calibrated separately.

References

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley, 1968.

Forbes, D. W., & Ingebo, G. S. An empirical test of the content homogeneity assumption involved in Rasch item calibration. Paper presented at the Annual Meeting of the American Educational Research Association, Washington, DC, April 1975.

Hambleton, R. K. An empirical investigation of the Rasch test theory model. Unpublished doctoral dissertation, University of Toronto, 1969.

Hambleton, R. K., Swaminathan, H., Cook, L., Eignor, D., & Gifford, J. Developments in latent trait theory: A review of models, technical issues, and applications. Paper presented at the Annual Meeting of the American Educational Research Association, New York, 1977.

Kifer, E., & Bramble, W. The calibration of a criterion-referenced test. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, 1974.

Lord, F. M. A theory of test scores. Psychometric Monographs, 1952, 7.

Lord, F. M. The "ability" scale in item characteristic curve theory. Psychometrika, 1975, 40, 205-218.

Lord, F. M., & Novick, M. R. Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley, 1968.

Reckase, M. D. Development and application of a multivariate logistic latent trait model. Unpublished doctoral dissertation, Syracuse University, 1972.

Reckase, M. D. Ability estimation and item calibration using the one- and three-parameter logistic models: A comparative study. (Research Report 77-1). Columbia, MO: University of Missouri, Department of Educational Psychology, in preparation.

Rentz, R. R., & Bashaw, W. L. Paper presented at the Harcourt Brace Jovanovich Conference on the Rasch model, New York, 1971.

Ryan, J. P., & Hamm, D. W. Practical procedures for increasing the reliability of classroom tests by using the Rasch model. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, 1976.

Whitely, S. E., & Dawis, R. V. The nature of objectivity with the Rasch model. Journal of Educational Measurement, 1974, 11, 163-178.

Wood, R. L., Wingersky, M. S., & Lord, F. M. LOGIST: A program for estimating examinee ability and item characteristic curve parameters (Research Memorandum RM-76-6). Educational Testing Service, Princeton, NJ: June 1976.

Wright, B. D., & Panchapakesan, N. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969, 29, 23-48.

A COMPARISON OF CONVENTIONAL AND COMPUTER-BASED ACHIEVEMENT TESTING

ISAAC I. BEJAR
UNIVERSITY OF MINNESOTA

There are at least two types of adaptive achievement testing. In the first type, if there is little variability in achievement at testing time and the purpose of measurement is to classify students into one of two categories (e.g., in mastery testing), then it becomes profitable to adapt the length of the test to the individual. If, on the other hand, there is a fair amount of variability in achievement and the purpose of testing is to assess level of achievement, then in the second type it becomes profitable to adapt both the length *and* difficulty of the test to each individual.

Ferguson (1969) implemented an achievement testing system for the first type which, through the use of decision theory, tested an individual until a decision could be made that the student had reached a prespecified level of achievement. No one has implemented a computer-based achievement testing system for the second type, although there has been an increasing amount of research on computer-based ability measurement.

In implementing an adaptive achievement testing system, serious consideration should be given to the kind of test theory that will serve as the psychometric framework. This is important because it has implications for the creation and calibration of the item pool. The present study is based on item characteristic curve (ICC) theory, which seems to be a more flexible theory compared to theories that could be used in the two types of achievement testing situations described above. Accordingly, this paper is concerned with two questions: (1) Is it possible to construct an item pool for achievement testing based on ICC theory--that is, do ICC models fit observed achievement test data? (2) If the answer is positive, what is the efficiency of an adaptive achievement test compared to a typical paper-and-pencil classroom examination?

The study was done with the cooperation of the Biological Sciences Department at the University of Minnesota. An introductory course was chosen because it had the largest enrollment. The course is offered every quarter, and enrollment ranges from 1,000 to 1,500 per quarter. It is open to all students; both majors and non-majors in the natural sciences enroll. Students are, for the most part, freshmen; but a substantial number of sophomores also enroll. The sexes are about equally represented. According to the course staff, there seem to be no changes in the demographic characteristics of the students from quarter to quarter. Instruction is by means of videotaped lectures shown on closed circuit television and a compulsory laboratory. The course is divided into three units. The first unit covers Chemistry, Energy, and The Cell, in that order; the second unit covers Heredity and Reproduction; and the third unit covers Ecology and Evolution. At the end of each of the first two units a midquarter examination is given. The final examination not only covers these units, but in addition covers the third unit. This paper will be concerned primarily with the first unit.

The Item Pool

Most research to date on adaptive testing has been based on verbal ability (e.g., Lord, 1975; Vale & Weiss, 1975). Typically, verbal items are homogeneous in content; therefore, one of the most important assumptions of latent trait theory, unidimensionality, is justified. By contrast, in achievement testing items are not homogeneous in content. The question arises whether or not a unidimensional model is adequate. There is, of course, no general answer to the question; it must be investigated in every new setting.

Development of the Item Pool

The raw data from which the item pool was formed consisted of answer sheets for the first midquarter exam from a previous academic year. The data matrix for each quarter consisted of between 1,000 and 1,500 respondents and 55 items. Data for each quarter were analyzed separately by Urry's ESTEM program (see Urry, 1976) for estimating item parameters using a minimum chi-square criterion. Table 1 shows the number of items originally available in each of the content areas as well as the number rejected by the program. The program rejects an item if it cannot find a reasonable estimate of one of the three item parameters. This occured with 22% of all items, as seen in Table 1. The remaining 78% of the items presumably fit the three-parameter response model. However, the fact that it was possible to obtain parameters is not in itself evidence of fit. Since the items were from several content areas, it was decided to investigate fit more closely.

Table 1
Summary of Calibration Study

	Chemistry	Cell	Energy	Total
Unique items available[1]	53	60	33	146
Items rejected	16	13	3	32
Percent of items rejected	30	22	9	22
Items calibrated	37	47	30	114

[1] Includes items administered at the final exam.

Dimensionality of the Pool

Correlation of item parameter estimates. It was reasoned that if there were any departures from unidimensionality, they would probably result from content area specific effects. To determine whether or not this was true, for each item a new set of parameter estimates was obtained. These were derived by including only items which belonged to a given content area. The rationale for this method was that if the different content areas measured a single dimension, the item characteristic curve estimated within a content area would be interchangeable with that derived for the entire examination. That is, the regression of content-area-derived parameters on parameters derived from the total examination would have a slope of 1.0 and an intercept of zero.

This was not found to be the case for either the *a* parameter or the *c* parameter. (This may have been attributable to the fact that there were relatively few items in each content area; it is known from simulation studies that a large number of items is required to obtain reliable estimates of the item parameters.) The results for the *b* parameter, on the other hand, showed a great deal of stability.

Table 2 shows the regression statistics for the regression of content-area-based *b* estimates on total test-based estimates for the two first mid-quarter exams. The prediction that the slope would be 1.00 and intercept would be zero was shown to be justified when taking the standard error into account for the first content area in both tests. For the second content area the slope was still 1.00, but the intercept no longer was zero in either midquarter. Finally, for the third content area neither the slope nor the intercept were what they should be. This replicable trend suggests that the metric based on the entire test is not interchangeable with the metric defined within content areas. That is, there is a unique component associated with each content area which was ignored when calibrating all items at once.

Table 2
Regression Statistics for the Regression of the
Content-Area-Based *b* Estimates on the Total Test-Based Estimates

	Slope		Intercept		
	Slope	S.E.	Int.	S.E.	Correlation
Winter					
Chemistry	.94	.03	.00	.03	.99
Cell	1.08	.06	-.41	.09	.98
Energy	.73	.08	.46	.13	.95
Spring					
Chemistry	1.03	.07	.16	.08	.98
Cell	.93	.04	-.31	.06	.99
Energy	.72	.12	.11	.20	.91

Correlation of achievement estimates. The next question was concerned with the extent of the content effect. To answer this question each of the three content areas was scored using the two sets of parameters. It was originally planned to use maximum likelihood scoring but difficulties occurred because there were a large number of cases which did not converge, presumably because some of the content areas had relatively few items. As a result, the data were scored using Owen's (1975) sequential procedure. The inter-content area correlations are seen in Table 3. Both of these matrices could be fitted perfectly by a one-factor model. The maximum likelihood estimates of the loadings on this factor and unique variances are also shown in Table 3.

Table 3
Observed Correlations Among Content Area Scores Using
Content-Area- and Test-Based Item Parameter Estimates

Content	Chemistry	Cell	Energy	λ	Uniqueness
Content-Area-Based					
Chemistry	.856[1]			.836	.29
Cell	.607	.895		.726	.47
Energy	.511	.444	1.052	.611	.63
Total Test-Based					
Chemistry	.836[1]			.834	.29
Cell	.623	.937		.747	.44
Energy	.594	.532	.798	.712	.50
Absolute Difference Between Correlations, Loadings, and Uniquenesses					
Chemistry	.020			.002	.00
Cell	.016	.042		.021	.03
Energy	.083	.088	.254	.101	.13

[1] Standard deviations are on the diagonal.

From Table 3 the importance of the content effect can be deduced by computing the difference in unique variances in the two solutions. As seen in Table 3, the estimated unique variances were the same or larger for the content-area-based solution. This is consistent with the earlier hypothesis that there is a unique component associated with performance on each content area beyond that accounted for by general achievement. These differences in unique variance are the proportion of variance attributable to the content area component. In the content areas of Chemistry and Energy this variance was negligible; in the content area of The Cell the variance was not negligible.

Conclusions. These results suggest that in calibrating an achievement test item pool, attention should be given to potential content area influences. It should be pointed out that factor analysis of inter-item correlations is not likely to provide assistance. Such a factor analysis was run and, although a predominant single factor was found, there was no detectable trace of content factors. The regression analysis previously reported appears to be much more powerful. Its usefulness, however, is limited because the analysis can be done only if subsets of items can be identified beforehand as belonging together.

Comparison of the Adaptive and Conventional Tests

Despite the presence of content area effects, each content area could not be calibrated separately to form separate item pools, since there were not enough items available. In effect, the presence of content effects was ignored; and the adaptive testing of achievement proceeded. Although this probably introduced some bias into the results, these scores would not be

Table 4
Values of the *a*, *b*, and *c* Parameters for Items Used in the Conventional Test

Item	*a*	*b*	*c*
3060	.86	-1.31	.29
3067	1.07	-.76	.21
3065	1.17	-1.66	.35
3056	.71	.89	.26
3063	.91	1.51	.35
3073	1.43	-1.57	.31
3058	1.05	-.43	.35
3274	.85	-1.05	.26
3271	.95	1.32	.30
3055	1.71	-.65	.24
3072	1.02	.65	.32
3057	1.20	-1.35	.26
3064	.94	.86	.24
3069	.88	-.01	.35
3054	1.29	-.93	.31
3066	1.05	.53	.31
3268	.97	-.28	.18
3267	1.02	-1.22	.23
3272	1.06	-.81	.35
3070	.95	-1.28	.22
3008	.96	-1.75	.18
3018	1.31	.29	.29
3062	1.47	.43	.30
3061	.95	1.57	.30
3262	.81	.47	.35
3263	.99	2.29	.35
3447	1.18	.93	.32
3443	1.07	-1.64	.35
3438	.70	.21	.27
3448	1.40	.73	.30
3435	.83	-.61	.35
3439	1.36	.64	.32
3436	1.12	1.59	.35
3449	.91	1.26	.14
3440	1.52	2.00	.30
3437	1.95	.66	.28
3427	.92	1.51	.26
3445	1.19	.44	.34
3444	.88	.78	.35

used for grading purposes; therefore, the bias could not affect the individual personally. Comparing modes of administration is often difficult because of the inherent differences of the two testing procedures (cf. Sympson, 1975). Nevertheless, this is a question that must be faced. This study compared the first midquarter Biology examination covering the areas of Chemistry, Energy, and The Cell and a stradaptive test covering the same content areas. The data were collected during fall quarter 1976 and are independent of the data used in the item calibration.

Tests

Classroom test. The classroom exam consisted of the 39 items for which item parameter estimates were available out of the 55 items in the actual test. These 39 items had a mean discrimination of 1.09. The distribution of difficulty was slightly peaked. (It should be pointed out that the criterion used by the Biology staff for assembling the test was a mixture of psychometric, pedagogical, and content considerations.) The item parameters for the 39 items are seen in Table 4.

The stradaptive test. The four major ingredients of the stradaptive strategy are the *item pool, entry point, branching rule,* and *termination criterion* (Weiss, 1973).

The item pool consisted of the 114 items described earlier (see Table 1). The items were assigned to one of nine strata in such a way that there were approximately the same number of items in each stratum. Within stratum the items were placed so that although the content areas were alternated, the most discriminating items were at the top of the stratum. The stradaptive item pool is seen in Table 5.

The entry point for the stradaptive test was determined by the students' reported GPA. For example, if the student reported a GPA of 3.75 or higher, the entry point was at the ninth stratum. At the other extreme, if the student's GPA was less than 2.00, the entry point was the first stratum (i.e., the easiest stratum).

The branching rule used in the present study was to present an item from the next more difficult stratum following a correct answer and an item from the next less difficult stratum following an incorrect answer. After responding to the first item in the entry stratum, the student was given the first item from the next lower stratum if the answer was incorrect or the first item from the next higher stratum if the answer was correct. Thereafter, the student was branched to the next unadministered item in the next higher or lower stratum, depending on whether the answer was correct or incorrect. The exception to this rule occurred if the testee was at the most difficult stratum. In that case, after a correct answer the next item in that same stratum was given. Similarly, for the student in the least difficult stratum, an incorrect response led to the next item in that stratum.

Table 5
Values of the *a*, *b*, and *c* Parameters for Items in the Stradaptive Test by Stratum

Item	*a*	*b*	*c*
Stratum 9:(most difficult):			
3209	2.77	2.29	.29
3417	2.67	3.02	.35
3033	1.54	2.44	.35
3251	2.60	2.39	.35
3406	1.31	2.48	.35
3045	1.02	2.48	.29
3242	.94	2.40	.35
3407	1.02	2.41	.29
3241	.91	2.09	.13
3414	.88	2.29	.32
3402	.83	2.44	.35
3247	.82	2.42	.35
3228	.67	2.49	.31
Stratum 8:			
3409	4.68	1.28	.00
3234	3.54	1.73	.00
3018	.89	1.25	.35
3204	1.14	1.66	.35
3422	1.47	1.50	.35
3411	1.36	1.23	.35
3250	.91	1.94	.29
3206	.74	1.51	.21
3410	1.30	1.34	.31
3429	1.25	1.24	.28
3419	1.23	1.48	.25
3421	1.17	1.15	.35
3427	.92	1.51	.26
3420	.68	1.62	.35
Stratum 7:			
3408	2.51	1.05	.31
3258	1.24	.81	.35
3432	1.72	.67	.35
3048	1.35	.66	.33
3413	1.40	.76	.35
3219	1.23	.62	.21
3035	.90	.68	.28
3433	1.35	.86	.30
3230	.90	.87	.35
3012	.75	.80	.38
3260	.71	.84	.28
Stratum 6:			
3047	1.66	.44	.29
3213	.93	.52	.35
3041	1.51	.23	.35
3405	1.40	.55	.32
3218	.82	.58	.12
3019	1.31	.29	.29
3207	.70	.46	.28
3431	.70	.28	.20
3000	1.24	.52	.35
3046	1.18	.24	.22
3042	1.15	.37	.27
3050	1.13	.35	.18
3034	1.01	.37	.28
Stratum 5:			
3220	1.79	-.03	.26
3005	1.43	.11	.35
3425	1.36	.17	.23
3039	1.12	.12	.34
3214	1.12	.03	.23
3412	1.12	.19	.35
3051	1.29	.21	.28
3403	.99	.18	.19
3211	.88	.01	.13
3002	.82	.13	.14
3426	.68	.07	.22
3423	.66	.16	.27
Stratum 4:			
3256	2.31	-.33	.26
3430	1.15	-.30	.29
3031	1.47	-.33	.35
3254	2.28	-.17	.27
3237	1.54	-.37	.18
3404	.65	-.29	.35
3244	1.35	-.44	.23
3240	.98	-.28	.15
3208	.76	-.16	.12
3006	.77	-.37	.33
3259	.69	-.41	.20
Stratum 3:			
3021	1.96	-.49	.21
3217	1.06	-.48	.14
3038	1.71	-.93	.21
3215	1.59	-.82	.23
3011	1.32	-.86	.20
3216	1.27	-.62	.18
3221	1.25	-.52	.17
3049	1.15	-.71	.18
3255	1.14	-.72	.26
3246	1.10	-.72	.28
3022	1.01	-.48	.30
3017	.99	-.58	.16
3224	.80	-.50	.27
Stratum 2:			
3023	2.40	-1.15	.35
3202	1.81	-.99	.21
3415	.85	-.96	.35
3245	1.34	-.96	.21
3236	1.26	-1.20	.33
3020	1.23	-1.28	.17
3028	1.12	-1.26	.35
3226	1.09	-.98	.20
3210	1.04	-1.22	.35
3239	1.04	-1.13	.21
3013	1.00	-.97	.39
3257	.98	-1.02	.25
3036	.92	-1.18	.16
3014	.86	-1.24	.14
3238	.82	-1.06	.21
3032	.77	-1.06	.27
Stratum 1:			
3027	1.67	-1.38	.35
3249	.91	-1.69	.17
3428	.90	-1.56	.35
3205	1.25	-1.53	.19
3235	1.15	-1.40	.28
3029	1.13	-1.50	.28
3201	1.07	-1.34	.23
3008	.96	-1.75	.18
3252	.79	-1.77	.35
3003	.96	-1.76	.34
3044	.87	-1.42	.15

The termination rule used in this study was that testing was stopped if in any stratum a student had answered 5 items and 20% or more of them had been incorrect answers, or if 50 items had been administered, whichever occurred first.

Scoring. Both the adaptive and classroom data were scored by the method of maximum likelihood, using the Newton-Raphson numerical procedure with a set of locally written programs. For scoring purposes the item parameter estimates were edited so that the maximum discrimination was 2.50, the maximum absolute value of the difficulty parameter was set to 3.00, and the maximum "guessing" parameter was set to .35. Less than 1% of the ability estimates failed to converge during scoring.

Criteria for Comparison

One of the most important contributions of latent trait theory to psychometrics has been the concept of information. Unlike reliability and related concepts, information is a local measure of the accuracy of estimation of a testee's achievement levels.

Samejima (1969) defines the test information function, in general, as

$$I(\theta) = E \frac{\partial^2 L_v(\theta)}{\partial \theta^2} , \qquad [1]$$

where $L_v(\theta)$ is the log-likelihood function of the response vector v. Thus, test information is the expected value of the second derivative of the log-likelihood function. This apparently arbitrary quantity is useful because its reciprocal, $1/I(\theta)$, is the minimum sampling variance of an estimator. As such, it is a measure of the best that could be accomplished in estimating with a given response model if an appropriate scoring procedure is used. Maximum likelihood estimation provides, asymptotically, estimators with that property.

Since it is an expected value, information is, in a sense, a prediction of the model; and in fact, it does not depend on a response vector. This is a useful property when making theoretical comparisons. For empirical comparisons, however, it seems more appropriate to base the result on a statistic closer to the data. That statistic may be called the *observed* information (cf. Edwards, 1972). Samejima (1973) has referred to observed information as the *response vector information function*. Equations 2 and 3 give, respectively, the expressions for the test information function and the response vector information function.

$$I(\theta) = \sum_g D^2 a_g^2 \, \Psi[DL_g(\theta)] - P_g(\theta) D^2 a_g^2 \, \Psi[DL_g(\theta) - \log c_g] \qquad [2]$$

$$\hat{I}(\theta) = \sum_g D^2 a_g^2 \, \Psi[DL_g(\theta)] - u_g(\theta) D^2 a_g^2 \, \Psi[DL_g(\theta) - \log c_g] \qquad [3]$$

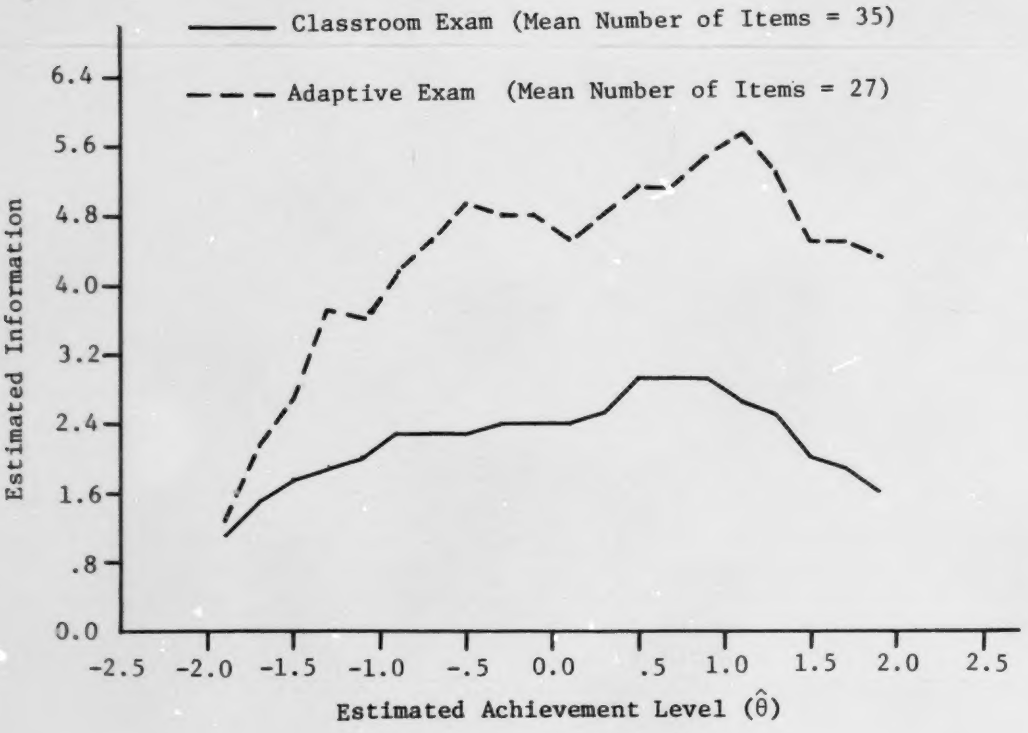
Equations 2 and 3 are identical with the exception that the second term on the right is weighted by $P(\theta)$ (i.e., the probability of passing the item) in one case and by u_g (=1 if the answer is right, 0 otherwise) in the other. To the extent that $P(\theta)$ is an accurate estimate of the probability that $u_g=1$, these two kinds of information functions will differ very little.

Results

Approximately 350 students came to the laboratory for testing between one day and three weeks after the classroom exam. For their participation, they received points toward their final course grade and a computer-printed report of the questions they had taken as part of the adaptive test.

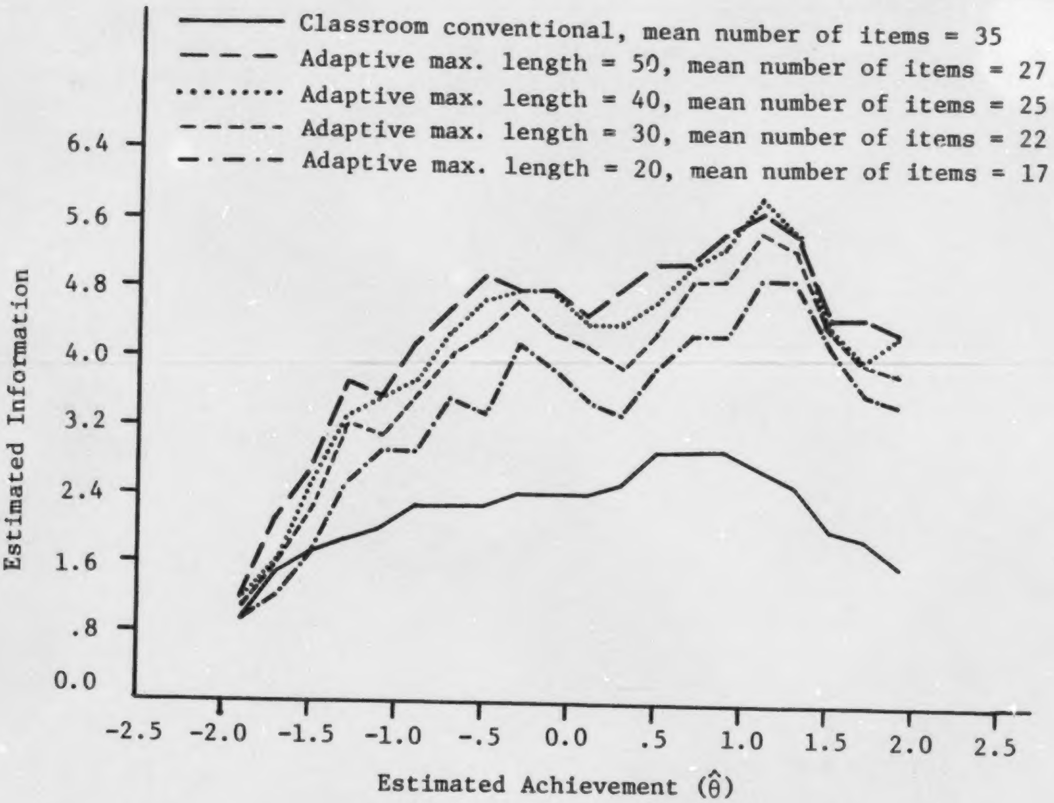
Adaptive vs. classroom test. Figure 1 shows the response vector information functions for the adaptive and classroom exams. To obtain the curves the maximum likelihood estimates of θ between -2.00 and +2.00 were divided into intervals of .20. The mean observed information for testees in a given interval was assigned to the midpoint of that interval. These values are plotted in Figure 1. The adaptive test is far superior; however, several factors must be considered to put this in perspective.

Figure 1
Response Vector Information Curves for
Classroom Conventional Test and Adaptive Test



As indicated earlier, the stradaptive test had variable test length. The mean test length in the 20 intervals of θ was fairly constant. The mean test length across all students was 27 items; for the conventional test the mean was 35. Despite the fact that the stradaptive test was shorter, on the average, it yielded more information. To see in retrospect what the results would have been with a shorter adaptive test, the adaptive data were rescored reducing the maximum test length from 50 to 40, 30 and 20 items. The mean number of items for the maximum lengths of 40, 30, and 20 were respectively 25, 22, and 17 items. The results are seen in Figure 2. Note that the adaptive test with a maximum test length of 20 still yielded a substantially higher amount of information compared to the classroom exam, while shortening the test considerably.

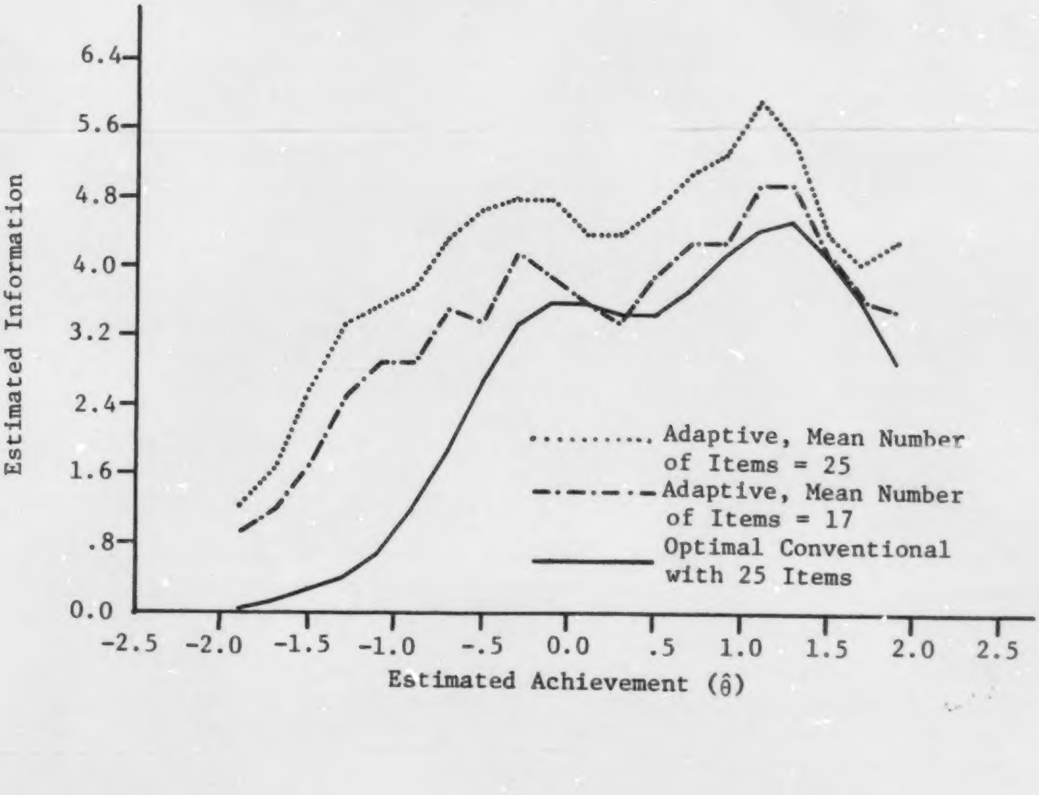
Figure 2
Response Vector Information Curves for
Classroom Conventional Test and Adaptive Test at Four Test Lengths



Adaptive vs. an ideal conventional test. These results are not surprising because the conventional test used for comparison was not designed to be optimal. A fairer comparison might have been to contrast the adaptive test against an ideal conventional test. For the ideal conventional test the top items were chosen from each of the seven most difficult strata until a maximum of 25 items was reached. This resulted in a conventional test with mean a=1.70 and approximately rectangularly distributed b's in the interval -1.00 to 3.00. The test information function of the conventional test and the response vector information function for the adaptive test with a maximum length of 20 and 40 items are seen in Figure 3.

The information for the optimal conventional test was very low for θ's below -1.00. This was a function of the distribution of item difficulties chosen for the test. It was reasoned that an optimal achievement test need not have very high information in the lower end of θ. In addition, by concentrating the difficulty of the items in a restricted range of θ, the conventional method was given a better chance against the adaptive test. The maximum information of this optimal conventional test was 4.59 at θ=1.10. The adaptive information functions also peaked at θ=1.10; and the shortest adaptive test yielded 10% more information with 30% fewer items on the average. Compared to the information for the optimal conventional test, the information for the adaptive test with a maximum length of 40 items (which resulted in a mean of 25 items per student, i.e., the same number of items as in the optimal conventional test) resulted in an even larger increase in information.

Figure 3
Information Curves for Optimal Conventional Test
and Adaptive Test at Two Test Lengths



Summary and Conclusions

Two questions have been addressed in this paper. The first question was whether it is possible to construct an item pool based on the ICC model which could be used in adaptive achievement testing. In general, the answer was found to be positive. The spread of difficulties and discrimination in the item pool were such that adaptive testing would be effective. However, unique response components associated with the different content areas were also identified. The magnitude of these components was not large, but must have introduced certain biases into the comparison of the two testing procedures. As a result of ignoring the influence of content-specific factors on test performance, the parameters of the estimated item characteristic curves derived from the entire test may have included a mis-specification bias. Since the adaptive testing procedure relied on these possibly misestimated item parameters for the sequential selection of items, it is likely that the advantage of adaptive testing over conventional achievement testing was underestimated.

The second question addressed was how effective adaptive achievement testing is compared to conventional testing. The answer was that adaptive testing can drastically reduce testing time while yielding more precise scores than an actual conventional or an ideal conventional test. Although the answer is gratifying, it is one which should be expected from theoretical studies. In fact, the stage is rapidly being approached in which the increased efficiency of adaptive testing is no longer an issue. Future research in adaptive achievement testing should concern itself with the truly unique needs of achievement testing. Two such needs are the ability to perform multi-content branching and the need to assess growth. Work on multi-content branching is already under way (Urry, 1977; Weiss & Brown, 1977). Little, however, seems to have been done in the areas of the assessment of growth by means of computerized testing. Hopefully, that gap will be filled in the near future.

References

Brown, J., & Weiss, D. J. An adaptive testing strategy for achievement test batteries (Research Report 77-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1977.

Edwards, A. W. F. Likelihood: An account of the statistical concept of likelihood and its application to scientific inference. London: Cambridge University Press, 1972.

Ferguson, N. L. The development, implementation, and evaluation of a computer-assisted branched test for a program of individually prescribed instruction. Unpublished doctoral dissertation, University of Pittsburgh, 1969.

Lord, F. M. A broad-range tailored test of verbal ability. In C. L. Clark (Ed.), Proceedings of the First Conference on Computerized Adaptive Testing. Washington, DC: U.S. Civil Service Commission, Bureau of Policies and Standards, 1975.

Owen, R. J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of the American Statistical Association, 1975, 70, 351-356.

Samejima, F. Estimation of latent ability using a response pattern of graded scores. Psychometrika, 1969, Monograph Supplement No. 17.

Samejima, F. A comment on Birnbaum's three-parameter logistic model in the latent trait theory. Psychometrika, 1973, 38, 221-234.

Sympson, J. B. Evaluating the results of computerized adaptive testing. In D. J. Weiss (Ed.), Computerized adaptive trait measurement: Problems and prospects (Research Report 75-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1975. (NTIS No. AD A018675)

Urry, V. W. A five-year quest: Is computerized adaptive testing feasible? In C. L. Clark (Ed.), Proceedings of the first conference on computerized adaptive testing. Washington, DC: U.S. Civil Service Commission, Bureau of Policy and Standards, 1976.

Urry, V. W. A multivariate model-sampling procedure and a method of multidimensional tailored testing. Paper presented at Computerized Adaptive Testing '77 Conference, Minneapolis, July, 1977.

Vale, C. D., & Weiss, D. J. A study of computer-administered stradaptive ability testing (Research Report 75-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1975. (NTIS No. AD A013185)

Weiss, D. J. The stratified adaptive computerized ability test (Research Report 73-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1973. (NTIS No. AD 768376)

Acknowledgments

This research was supported by funds from the Navy Personnel Research and Development Center, Army Research Institute, Air Force Human Resources Laboratory, Defense Advanced Research Projects Agency, and the Office of Naval Research, monitored under contract No. N00014-76-C-0627 NR150-389 with the Office of Naval Research. The assistance of Kathleen A. Gialluca and G. Gage Kingsbury in data analysis is deeply appreciated, as is the cooperation of Kathy Swart and Norman Kerr of the General Biology staff in the College of Biological Sciences at the University of Minnesota, in providing access to the students and data.

DISCUSSION: SESSION 7

HARIHARAN SWAMINATHAN
UNIVERSITY OF MASSACHUSETTS

It is always easy to criticize the presenters in an Epimethean sense, i.e., in hindsight. Since I can only point out the problems with the papers in the time I have, I hope that the importance of these papers does not get diminished.

I should like to begin by indicating my agreement and, mainly, my disagreement with McBride's paper. One of the issues pertains to Hambleton's statement about the basic problem in the area of criterion-referenced testing (CRT). In order to integrate research in this area, it is important to work with accepted terminology rather than creating new terminology and symbols. We are finally coming to a closure concerning what is meant by criterion-referenced measurement in terms of domain and uses of test scores. It is unfortunate that McBride did not build upon that. He created some new terms, such as task and domain of task, which are applicable to the research he is doing for the Army; however, more care should be taken with the terms used in order to communicate with other researchers.

I also take issue with McBride's lack of incorporation of research or ideas that have already been expounded in this area. For example, in terms of the application of latent trait theory to criterion-referenced testing, there have been at least several suggestions regarding what should be done and how it should be done. Hambleton and Cook (1977) devote an entire section to the application of latent trait theory to CRT. All the issues were also discussed in a paper by Hambleton, Swaminathan, Cook, Eignor and Gifford (1977). These issues have also been raised by other researchers, for example, Huynh (1976), who incorporated ideas stemming from latent trait theory in making decisions and estimating mastery scores in criterion-referenced testing. McBride does not refer to these papers nor build upon the ideas that have already been expounded.

At this point in time, we should not be describing the applicability of latent trait theory to criterion-referenced testing; it has already been done. What should be done instead is to address some of the more important issues in the application of latent trait theory to criterion-referenced testing. For example, one area which needs to be addressed is specification of the model that is appropriate in the context of criterion-referenced testing. Is there any advantage in specifying one-parameter, two-parameter, or three-parameter models in latent trait theory? and what are the consequences? This is a critical issue, since one of the big problems with latent trait models is the estimation of parameters. In a CRT there are typically four or five items. The question is, can a three-parameter logistic model be specified and the item parameters as well as the ability parameters be estimated without running into the well-known estimation problems?

It would be futile to estimate the item parameters over and over again when there is a domain of items. What could be done is to calibrate the items beforehand; to estimate the *a*'s, the *b*'s, and the *c*'s; and then when actually dealing with a CRT, just to estimate the student's achievement scores or ability. Samejima (1972) has shown that in the case of known item parameters, maximum likelihood estimators for ability parameters can be found readily. However, the problem of specifying the correct latent trait model still remains to be addressed.

The other problem is, exactly how is latent trait theory utilized in criterion-referenced situations? McBride pointed out the two main uses: The first is classifying individuals into mastery states; and the second, which was discussed earlier, is estimating the ability parameters. Classifying individuals into mastery categories requires the use of the test score metric. The test score, in this case, is the proportion of items correct. When making mastery decisions, a cut-off domain score has to be specified and converted to the underlying theta metric via the test characteristic curve (Hambleton, 1977). In order to obtain the test characteristic curve, *all* the items in the domain must be calibrated; this may turn out to be an extremely difficult task. Considerable further work is clearly needed in this area.

The notion of time-referenced tests that McBride introduced seems to be an interesting idea. At least statistically it appears to be a valid technique because there is now an *observable* metric--time. The important questions raised are, what are the implications? what is the use of time-referenced tests? will this relate the performance of an individual to the time required for an individual to complete an item? The measurement of time is bound to introduce an added problem to the already problem-laden field of measurement. In addition, what would happen to the notion of local independence? Finally, if we want to relate scores obtained in a time-referenced test to an underlying ability continuum, we get back into the realm of latent trait theory; however, the problem is even more complex now, since we have a two-stage model. I would like to speculate about what could happen in terms of time dependent tests. Perhaps we may eventually run into something like Heisenberg's uncertainty principle, and it may be impossible to locate an individual in terms of time and his/her ability.

Applying latent trait theory to CRT is a viable approach. It is an important application, and it solves a great number of problems, especially in terms of domains that have varying item difficulties. McBride assumed that all the items have the same item characteristic curve, which is an overly restrictive assumption. If the latent trait model is to be used, that assumption can be discarded; otherwise, not much will be gained by using latent trait theory. Item difficulties can be permitted to differ, the one-parameter logistic model can be used, and item parameters and ability can be estimated. Domains and CRTs can be expected to be unidimensional, since domains having that property can be constructed. The latent trait model thus has tremendous implications. The important questions are, how many items are needed to estimate the parameters? can parameters or ability be estimated using four items? or must other methods be resorted to? One approach is to use Bayesian methods, which would incorporate prior information and, hence, be valid in criterion-referenced testing situations. These are the kinds of issues I would like to see discussed, at least in the next conference.

If latent trait models are to be successfully employed, the assumptions underlying latent trait models have to be studied carefully. Of particular importance is the assumption of unidimensionality and the effect it has in the scaling of multidimenisonal tests. Hence, Reckase's paper is a paper that should be of considerable interest to practitioners (and for that matter, to theoreticians as well).

I did not understand some aspects of Reckase's paper. The first problem is that the logic of carrying out a principal components analysis and a principal factor analysis with phi coefficients, and yet only a principal components analysis based on the tetrachoric correlations, is not made clear. In order to obtain comparable results, comparable analyses should be carried out. The second problem concerns the results presented in Tables 3, 4, 5, and 6. These tables seem incomplete, since all the relevant data are not reported.

The discussion of the results appears to be equally perplexing. In discussing the results of Table 4, Reckase pointed out that the 3 PL discrimination estimates correlated highly with the second factor (r=.97). Hence, he concluded that the 3PL model "was measuring the second factor." However, the correlation between the discrimination estimate and the first factor was equally high (-.96). This information is completely omitted in interpreting the relationship between the discrimination estimates and the factors. Another problem with the results in Table 4 is the finding that the "1 PL chi-square probability of fit value did not correlate significantly with any of the other statistics." This is correct, but the conclusion that "this indicates that neither variation in discrimination nor the multifactor nature of the data can be used to explain the lack of fit for this data set" seems to be unwarranted. There is no reason for the goodness of fit statistic to correlate with the other statistics. The goodness of fit statistic is a non-linear function of the item difficulty in the 2 PL model; hence, it should not correlate with the estimate of item difficulty. As Reckase points out later, item difficulty and item discrimination in the 3 PL model are not independent; therefore, the goodness of fit statistic can be expected to correlate poorly with the estimates of the discrimination index as well. This is clearly borne out by the results in Table 7.

The interpretations of the results presented in Table 5 are also questionable. Correlations between ability estimates, raw scores, and factor scores are reported in this table. It is not clear what the column headed θ represents. Probably this represents the correlation between the ability estimate from the one-parameter model and the true ability, θ. It would be interesting to compare this with the correlation between the ability estimate from the 3 PL model and θ, but this is not reported. The interpretation of the results for data set 250 AN appears to be relatively straightforward. However, even in this case there are some ambiguities. The statement "although the 3 PL ability estimates were clearly more closely related to the second rotated factor than the first. . ." appears to be based on the results given in Table 4 and, as pointed out earlier, is clearly not correct. Reckase does point out correctly in this connection that the "3 PL model was estimating the second factor (though rather poorly). . .", but somehow the *poorness* of the relationship is ignored in later interpretations.

The interpretation of the results for data set 950 A9 is misleading. It is pointed out that (re Table 5) the "2 PL ability estimates correlated highly with the raw scores and had moderate correlations with the principal components and varimax scores." The varimax scores were not reported for this data set. In addition, the correlations between the 1 PL ability estimate and raw scores was .98, while the "moderate" correlation between the 1 PL ability estimate and the component scores were .98 and .98, respectively.

The most puzzling aspect of this phase of the analysis is that after it is pointed out that "in all cases except two, the results show that the two models were measuring the first factor. . .", Reckase concludes that "these results show that the 3 PL model estimates one factor when independent factors are present, while the 1 PL model estimates the sum of the factors." The data clearly do not warrant this conclusion.

In examining the relationship between the trait that is measured and the "dominancy" of the first factor, Reckase has obtained some interesting results. Again, some statistics that could be of interest are omitted. For instance, Reckase demonstrates that the standard deviation of the difficulty estimates increased as the size of the largest eigenvalue decreased. The conclusion is that the 3 PL estimation procedure becomes less stable when the test is factorially complex. Does a similar result hold for the discrimination index also? This is a more interesting result than the reported result, which shows that average discrimination increases with the size of the eigenvalue. In connection with this, it should be pointed out that, in general, a large eigenvalue does not result in a high discrimination. Although this implication is not intended in the paper, it should be pointed out that in some sense the contrary is true. If the items have high discrimination indices, and if these differ from each other, then the matrix of correlations based on such items will not yield a single factor.

The relationships between factor analysis and latent trait theory are extremely complex and are indeed difficult to study, particularly since the two procedures are not strictly comparable. Reckase's study has shown that some interesting possibilities exist, although the results reported in this study do not greatly support the conclusions drawn.

Clearly, considerable further research is needed in order to establish relationships, if any, between factor models and latent trait models. One possible suggestion for further research is to study the lack of fit through carefully simulated data with varying numbers of factors and to correlate the fit statistic with an index of factorial simplicity (or complexity). This relationship should be studied for the various latent trait models. Another suggestion is to approach the problem of dimensionality from the direction opposite to that employed by Reckase. This would involve generating data to fit the latent trait models and carrying out a factor analysis on the data by varying the parameters involved to study the relationship between the two models. The theory developed by McDonald (1967) can be employed for this purpose.

The last paper I want to discuss briefly is Bejar's paper. It is a perfectly valid and well-done study, comparing conventional and computer-based achievement testing; and it has a number of implications. However, there are a couple of comments he has made which disturb me. First, I think Bejar protested too much in terms of the underlying content level, which was not homogeneous. In order to determine if differences due to content exist, he did a factor analysis study, looked at the factor loadings, and concluded that a content effect was present. He then looked at the differences between correlations and raised them to the fourth power. Obviously, when correlations are raised to the fourth power, an even smaller number will be obtained. After obtaining the small number, Bejar dismissed the content effect. In other words, after going through all the trouble of establishing a content effect, he then stated that it was not important. He could have started with that assumption, and in some sense it would have been justified.

The second comment Bejar made that disturbs me was in terms of the information function. The information function, of course, is a wonderful thing to use. However, it is not specific to latent trait theory; it is a property of maximum likelihood estimators. It is a well-known property of maximum likelihood estimators that they are asymptotically normally distributed. Furthermore, if maximum likelihood estimators exist, then they are asymptotically normally distributed with a variance-covariance matrix given by the inverse of the information matrix. The information matrix is thus the asymptotic variance of the maximum likelihood estimator.

I do not think it is really advisable to use that notion of information unless one is dealing with maximum likelihood estimates. If the notion of information is going to be used without using maximum likelihood estimates, then perhaps we should start with a different definition of information. If test efficiency is to be examined in terms of information and if it is not desirable to use maximum likelihood estimates, then a different rationale for using information functions must be found. For example, if a Bayesian procedure is employed, then the appropriate quantity is the variance of the posterior distribution. Finally, the information function is asymptotic; this should be kept in mind for very small sample sizes.

I enjoyed the papers in this session very much and found them to be extremely thought provoking. The authors should be congratulated on their efforts; they have addressed important and extremely difficult issues and have pointed the way to future research. It is obvious that great strides have been made in the area of latent trait theory which would have been impossible without research efforts such as those presented here.

References

Hambleton, R. K. Contributions to criterion-referenced test theory on the use of item characteristic curves and related concepts. Paper presented at the annual meeting of American Educational Research Association, 1977.

Hambleton, R. K., & Cook, L. L. Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 1977, 14, 75-96.

Hambleton, R. K., Swaminathan, H., Cook, L., Eignor, D., & Gifford, J. A. Developments in latent trait theory: A review of models, technical issues, and applications. Symposium presented at the annual meeting of the American Educational Research Association, 1977.

Huynh, H. Statistical consideration of mastery test scores. Psychometrika, 1976, 41, 65-78.

McDonald, R. P. Non-linear factor analysis. Psychometric Monograph, No. 15, 1967.

Samejima, F. A general model for free response data. Psychometric Monograph, No. 18, 1972.

SESSION 8

COMPUTER-BASED TESTING AS AN ALTERNATIVE TO PAPER-AND-PENCIL TESTING

RESEARCH ON COMPUTER-BASED PERCEPTUAL TESTING

DAVID R. HUNTER
AIR FORCE HUMAN RESOURCES LABORATORY

ADMINISTERING PAPER-AND-PENCIL TESTS BY COMPUTER, OR THE MEDIUM IS NOT ALWAYS THE MESSAGE

JANE SACHAR AND J. D. FLETCHER
NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER

DISCUSSION

DAVID J. WEISS
UNIVERSITY OF MINNESOTA

SESSION 8: ABSTRACTS

RESEARCH ON COMPUTER-BASED PERCEPTUAL TESTING

DAVID R. HUNTER

Tests of several psychomotor/perceptual abilities were implemented in a mini-computer-based testing system and administered to large samples of both Air Force officers and enlisted personnel. A number of these tests were evaluated through the use of factor analytic techniques to determine their relationship with paper-and-pencil measures. Additionally, the validities of some of the measures for the prediction of training outcomes in pilot and navigator training for officers and technical training courses for enlisted personnel were evaluted. In general, these psychomotor/perceptual tests seem to measure abilities untapped by existing Air Force selection instruments; and several of them may make significant contributions to the validity of selection procedures.

ADMINISTERING PAPER-AND-PENCIL TESTS BY COMPUTER, OR THE MEDIUM IS NOT ALWAYS THE MESSAGE

JANE SACHAR AND J. D. FLETCHER

Two empirical studies of computerized sequential tests are reported. In the first study, a counterbalanced design was used to present parallel versions of two tests to subjects who received one version of each test under computer administration and one version under paper-and-pencil administration. In the second study, a test-retest design was used in which subjects either received both versions under computer administration or both versions under paper-and-pencil administration. Results indicated (1) differences in mean test scores due to mode of administration; (2) differences in reliabilities of the tests under different modes for high and low ability subjects; (3) similar item characteristics under the two administration modes for high ability subjects, but different item results for low ability subjects; and (4) a different relationship between response latency and correctness for high and low ability subjects.

RESEARCH ON COMPUTER-BASED PERCEPTUAL TESTING

DAVID R. HUNTER
AIR FORCE HUMAN RESOURCES LABORATORY
BROOKS AIR FORCE BASE, TEXAS

Interest in the measurement of perceptual and psychomotor abilities for selection purposes in the Air Force extends back to the early years of World War II when the Psychological Research Units of the Army Air Corps were formed. Among the tests which these units developed were the Complex Coordination Test (sometimes called the Mashburn Test), a Two-Hand Rotary Pursuit Test, and a Discrimination-Reaction Time Test (Melton, 1947). These tests demonstrated considerable validity for the selection of air crew personnel, but were eventually abandoned in the early 1950s because of the difficulties encountered in the calibration and maintenance of the electro-mechanical devices.

As advanced technology of solid state devices became readily available in the late 1960s, Air Force interest was again focused on the use of measures of perceptual and psychomotor abilities--this time obtained from more reliable devices. The decision was made at that time to develop a computer-based testing system of sufficient flexibility to allow for the assessment of a variety of abilities through the addition of appropriate response manipulanda and suitable modification of the controlling computer programs. The alternative approach--that of developing special hard-wired devices for each test--was considered far less flexible than the use of computer-controlled testing for exploratory research of the type envisioned.

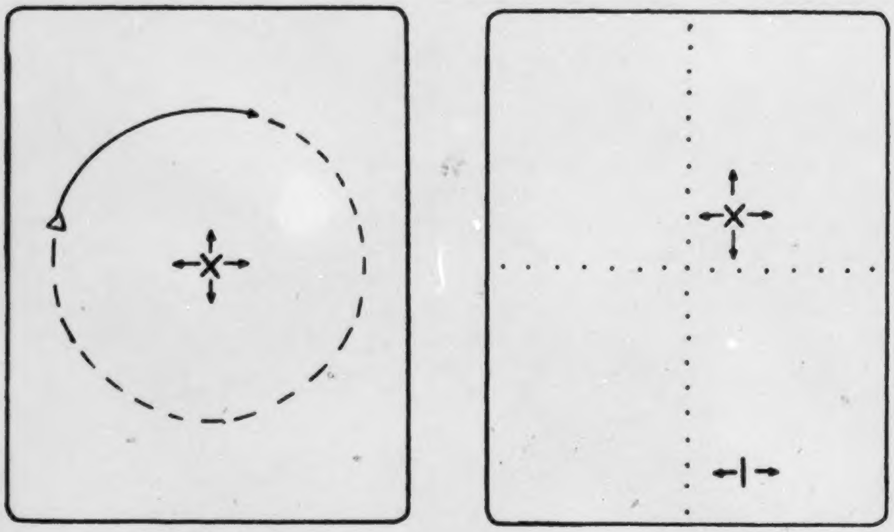
Test Batteries

Original Test Battery

Two-Hand Coordination. The first system developed consisted of a PDP-8/L mini-computer with only 4K memory, two cathode ray tube (CRT) displays, and two sets of joysticks. On this system were implemented two tests of psychomotor/perceptual ability, patterned somewhat after the World War II tests (Sanders, Valentine, & McGrevy, 1971). The first of these tests, Two-Hand Coordination, was a pursuit-tracking test in which the subject used two hand controllers to control the movement of an X-shaped cursor (see Figure 1). One hand controller moved the cursor left and right, while the other moved it up and down on the CRT. The target was a triangle which moved in a circular path on the CRT at a somewhat irregular speed, and the subject's task was to keep the cursor as close as possible to the target.

Displacements of the target from the cursor in both the *X*- and *Y*-axis were measured at a fixed time interval (approximately 60 times per second). These error scores were summed over five one-minute periods to produce the basic scores obtained from the test.

Figure 1
Stimuli used in the Two-Hand Coordination
Test and the Complex Coordination Test



TWO-HAND
COORDINATION
TEST

COMPLEX
COORDINATION
TEST

Complex Coordination. The second test, Complex Coordination, was a compensatory tracking task requiring the manipulation of a single joystick to control the movement of an X-shaped cursor, while using both feet to control a short vertical line displayed near the bottom of the CRT, as shown in Figure 1. The subject was required to hold both the cursor and the line at stationary points on the CRT which were identified by fixed intersecting lines. Error scores were obtained for the cursor in both the *X*- and *Y*-axis and for the vertical line along its dimension of movement. Additionally, a count was kept of the instances in which the subject lost control of the cursor and the computer had to reset the cursor to the center of the screen.

Validation. Both of these tests were validated against performance criteria for subjects undergoing Undergraduate Pilot Training (UPT) and were found to correlate significantly with UPT training criteria. Correlations between scores from the Two-Hand and Complex Coordination Tests and a UPT graduation/elimination criterion are given in Table 1 (McGrevy & Valentine, 1974).

Table 1
Validity Correlations (r) of Two Air Crew Psychomotor
Tests Against Pass/Fail Criterion

Two-Hand Coordination (N=121)		Complex Coordination (N=92)	
Error and Interval	r	Error and Interval	r
X Axis Error		*X* Axis Error	
Minute 1	-.24*	Minute 1	.04
Minute 2	-.19*	Minute 2	-.29*
Minute 3	-.07	Minute 3	-.22*
Minute 4	-.14	Minute 4	-.19
Minute 5	-.08	Minute 5	-.24*
Y Axis Error		*Y* Axis Error	
Minute 1	-.17	Minute 1	-.15
Minute 2	-.17	Minute 2	-.29*
Minute 3	-.04	Minute 3	-.21*
Minute 4	-.07	Minute 4	-.20*
Minute 5	-.04	Minute 5	-.27*
		Z Axis Error	
		Minute 1	-.17
		Minute 2	-.27*
		Minute 3	-.29*
		Minute 4	-.20*
		Minute 5	-.27*

*Significant beyond .05 level

Revised Test Battery

Hardware. Following the development of the two tests for pilot selection, the PDP-8 testing system was upgraded through the addition of a seven-track tape drive and an additional 8K of memory. Interface and subject testing stations were also upgraded to allow for the presentation of instruction via a computer-controlled cartridge tape player; and an expanded response panel was installed at the two test stations. An overview of this system is shown in Figure 2.

Tests. A seven-test battery was developed and implemented on the system. The seven tests in the battery were: (1) Kinesthetic Memory; (2) Perceptual Speed; (3) Performance under Stress; (4) Associative Learning; (5) Memory (Immediate/Delayed); (6) Concept Identification; and (7) Performance under Divided Attention. Figure 3 shows a sample of the geometric figures used in the Concept Identification Test. Scores obtained from these tests (Table 2) included the number of correct answers and response and perception latencies.

These tests were patterned, in part, on tests developed by McLaurin (1973) under contract to the Air Force Human Resources Laboratory. McLaurin's tests demonstrated validity for the selection of Air Force personnel for ground equipment courses; however, they were implemented on a hard-wired system of limited flexibility.

Figure 2
Diagram of Revised Testing System

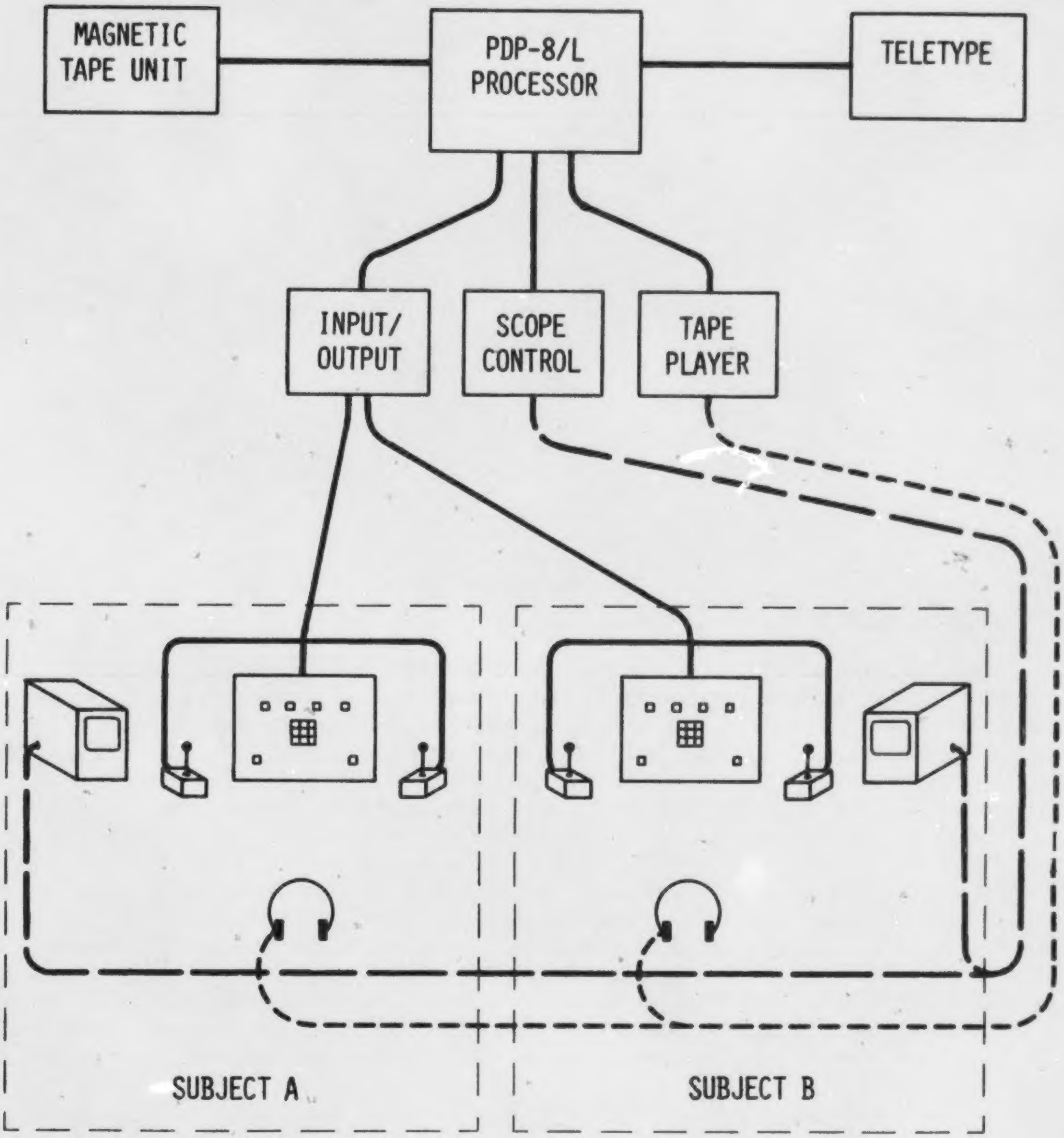

MAGNETIC TAPE UNIT
PDP-8/L PROCESSOR
TELETYPE
INPUT/ OUTPUT
SCOPE CONTROL
TAPE PLAYER
SUBJECT A
SUBJECT B

Figure 3
Geometric Figures Used in Test 6 -- Concept Identification

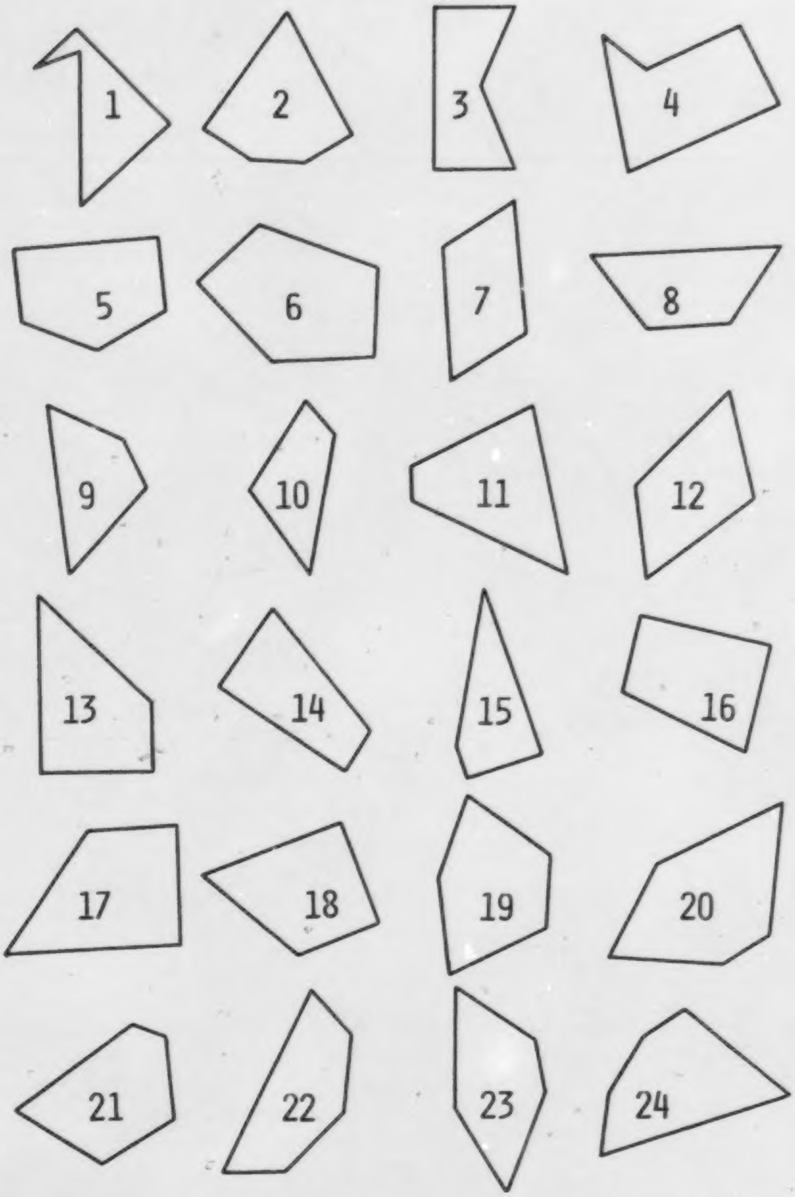

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24

Table 2
Scores Obtained from the Tests in the Perceptual/Psychomotor Battery

Test	Score
1. Kinesthetic Memory	Number of Correct Responses Response Time
2. Perceptual Speed	Number of Correct Responses Response Time Perception Time
3. Performance Under Stress	Number of Correct Responses Response Time Perception Time
4. Associative Learning	Number of Correct Responses
5. Memory (Immediate/Delayed)	Number of Correct Responses
6. Concept Identification	Number of Correct Responses
7. Performance Under Divided Attention	Summed Absolute Error

The battery of tests implemented on the PDP-8 testing system was administered to a large sample of enlisted personnel in a variety of career fields. At the same time, the subjects were given a four-hour paper-and-pencil battery consisting of the tests shown in Table 3.

Table 3
Paper-and-Pencil Reference Measures

Scale Reading
Letter Sets
Tool Functions
Electrical Information
Mechanical Principles
Word Knowledge
Word Grouping
Verbal Analogies
Block Counting
Point Distance
Electrical Maze
Pattern Detail
Rotated Blocks
Tools
Figure Analogies
Hidden Figures
Answer Sheet Marking
Table Reading
Large Tapping
Trace Tapping II
Discrimination-Reaction

Results. Analyses of the data indicated that the measures obtained from the psychomotor/perceptual battery were generally highly reliable (Hunter, 1975). Factor analyses (see Table 4) resulted in the identification of six factors that were specific to the psychomotor/perceptual battery, four factors that were specific to the paper-and-pencil measures, and one factor that was common to both batteries.

Subsequent validation of the battery for both enlisted personnel and navigator trainees showed that certain of the tests were valid predictors of training success. In particular, as Table 5 shows, the best single predictor of success in navigator training was the Kinesthetic Memory Test, with a

Table 4
Rotated Factors, Psychomotor/Perceptual Battery
(N=305)

Variable	Factor Loading	Variable	Factor Loading
Factor I (Visual Tracking)		**Factor II (Auditory Tracking)**	
Test 7 Performance Under Divided Attention-Line Error Minute 2	.93	Test 7 Performance Under Divided Attention-Tone Error Minute 2	.83
Test 7 Performance Under Divided Attention-Line Error Minute 3	.91	Test 7 Performance Under Divided Attention-Tone Error Minute 3	.82
Test 7 Performance Under Divided Attention-Line Error Minute 1	.89	Test 7 Performance Under Divided Attention-Tone Error Minute 4	.82
Test 7 Performance Under Divided Attention-Line Error Minute 4	.89	Test 7 Performance Under Divided Attention-Tone Error Minute 1	.81
Factor III (Figural Memory)		**Factor IV (Position Memory)**	
Test 5, Memory (Delayed) Part 2	.85	Test 2, Perceptual Speed-Correct Answers	.84
Test 5, Memory (Immediate) Part 1	.83	Test 1, Kinesthetic Memory-Correct Answers	.63
Test 3, Performance Under Stress-Correct Answers	.51		
Test 2, Perceptual Speed-Response Time	–.41	**Factor VI (Motor Speed)**	
Test 1, Kinesthetic Memory-Correct Answers	.36	Test 3, Performance Under Stress-Response Time	.90
Test 2, Perceptual Speed-Perception Time	–.36	Test 2, Perceptual Speed-Response Time	.71
Factor V (Associative Speed)		**Factor VII (Perceptual Speed)**	
Test 4, Associative Learning Part 1	–.69	Test 3, Performance Under Stress-Perception Time	.90
Test 1, Kinesthetic Memory-Response Time	.68	Test 2, Perceptual Speed-Perception Time	.70
Test 4, Associative Learning Part 2	–.54		
Test 6, Concept Identification-Correct Answers	–.45		
Test 2, Perceptual Speed-Perception Time	.32		

correlation of .32 between Mean Response Time and a dichotomous pass/fail criterion. Kinesthetic Memory was also one of the best predictors of training success in the enlisted sample.

Discussion

Because of the results summarized above, the Human Resources Laboratory is now in the process of acquiring a new, larger, and more sophisticated Computer-Assisted Measurement System. This system will be based upon the versatile and powerful PDP-11/34 mini-computer and will allow for the

Table 5
Validity Correlations for the Perceptual/Psychomotor Measures

Test and Score	Navigators* (*N*=77)	Enlisted** Personnel (*N*=395)
Kinesthetic Memory -- Correct Answers	.24	-.08
Kinesthetic Memory -- Response Time	-.32	-.22
Perceptual Speed -- Correct Answers	.05	-.19
Perceptual Speed -- Perception Time	-.16	-.10
Perceptual Speed -- Response Time	-.17	-.13
Performance Under Stress -- Correct Answers	.13	-.06
Performance Under Stress -- Perception Time	.05	-.05
Performance Under Stress -- Response Time	.01	-.09
Associative Learning (Part 1) -- Correct Answers	.07	.05
Associative Learning (Part 2) -- Correct Answers	.11	.01
Memory (Immediate) -- Correct Answers	.23	.08
Memory (Delayed) -- Correct Answers	.14	.05
Concept Identification -- Correct Answers	.11	.00
Performance Under Divided Attention -- Line Error, Min. 1	-.03	-.05
Performance Under Divided Attention -- Line Error, Min. 2	-.22	-.02
Performance Under Divided Attention -- Line Error, Min. 3	-.08	-.01
Performance Under Divided Attention -- Line Error, Min. 4	-.08	-.02
Performance Under Divided Attention -- Tone Error, Min. 1	-.07	.00
Performance Under Divided Attention -- Tone Error, Min. 2	-.01	-.01
Performance Under Divided Attention -- Tone Error, Min. 3	-.06	.02
Performance Under Divided Attention -- Tone Error, Min. 4	-.03	-.03

---

* $r_{crit.} = .22$, for $p < .05$

** $r_{crit.} = .10$, for $p < .05$

eventual expansion to a 20-station system. Each station will be equipped with a PDP-11/04 processor with 24K memory, a CRT display, two joysticks, a typewriter-like keyboard, and a special function keyboard.

It is expected that certain of the more promising tests from the preceding battery will be implemented on the new system, along with other tests of perceptual and psychomotor abilities. Additionally, this sytem will be used for the development and evaluation of adaptive testing procedures.

Because of their mutual dependence on sophisticated, typically computer-driven hardware, perceptual/psychomotor and adaptive testing procedures may be strongly linked in the future. The expenditures required for the operational application of either of these testing procedures may be such that neither procedure by itself can offer sufficient cost savings to warrant its use. When used in conjunction on the same equipment, however, the combined increase in testing effectiveness may offset the greater expense of these procedures.

The possible roles of perceptual/psychomotor testing, especially in conjunction with adaptive testing procedures, will be thoroughly investigated by the Human Resources Laboratory in the future. Hopefully, the investigation will indicate ways in which these procedures may be implemented for the improved selection of Air Force personnel.

References

Hunter, D. R. Development of an enlisted psychomotor/perceptual test battery (AFHRL-TR-75-60). Lackland Air Force Base, TX: Air Force Human Resources Laboratory, Personnel Research Division, November 1975.

McGrevy, D. F., & Valentine, L. D. Validation of two air crew psychomotor tests (AFHRL-TR-74-4). Lackland Air Force Base, TX: Personnel Research Division, Air Force Human Resources Laboratory, January 1974.

McLaurin, W. A. Validation of a battery of performance tests for prediction of aerospace ground equipment course grades (AFHRL-TR-73-20). Lackland Air Force Base, TX: Air Force Human Resources Laboratory, Personnel Research Division, November 1973.

Melton, A. W. (Ed.), Apparatus tests. Army Air Forces Aviation Psychology Program (Research Report No. 4), Washington DC: U.S. Government Printing Office, 1947.

Sanders, J. H., Valentine, L. D., & McGrevy, D. R. The development of equipment for psychomotor assessment (AFHRL-TR-71-40). Lackland Air Force Base, TX: Air Force Human Resources Laboratory, Personnel Research Division, July, 1971.

Acknowledgement

The views expressed herein are those of the author and do not necessarily reflect the views of the United States Air Force or the Department of Defense.

ADMINISTERING PAPER-AND-PENCIL TESTS BY COMPUTER, OR THE MEDIUM IS NOT ALWAYS THE MESSAGE

JANE D. SACHAR AND J. D. FLETCHER
NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER

Several attractive possibilities are offered by the use of computers to administer tests. Among these possibilities are reduction of costs in test administration and scoring, reduction in error variance in test scores, use of tailored testing in which the items are uniquely selected for each individual, and branched testing in which the tests themselves may be uniquely selected for each individual. Despite the high probable returns from investigations of computer-administered testing, recent surveys (e.g., Hansen, Johnson, Fagan, Tam, & Dick, 1973; Jensema, 1972) have indicated that little is understood about the implications of computer-administered testing for test reliability and validity. Moreover, investigations of computer-administered testing should usefully complement work in areas not directly concerned with testing, such as computer-assisted instruction, and in areas not directly concerned with computers, such as paper-and-pencil testing.

Much of the current research on computerized testing has focused on various strategies of tailored (Lord, 1970) or adaptive (Weiss & Betz, 1973) ability testing. These testing procedures "adapt" the test to the student on the basis of response patterns by presenting items successively more appropriate to the individual's ability level. This adaptive item selection enables the achievement of more accurate measurement.

Purpose

The purpose of this study is to better understand the implications of computer-administered testing by investigating the adaptation of standardized paper-and-pencil tests for computer administration. This study compared statistical characteristics of tests under paper-and-pencil administration with the same characteristics under computer administration.

Background

One difficulty in programming standard tests for computer administration lies in giving the student adequate review of the items answered. In paper-and-pencil test administrations, a student can review and change items already answered with relative ease; in computer test administrations, only one item may be displayed at one time. A compromise must be reached between the amount of flexibility a student has for reviewing his/her answers and the simplicity of his/her interactions with the computer. The simplest strategy is to allow students only one chance to answer each item on the test, with no opportunity to review responses. However, as Suppes, Fletcher, Zanotti, Lorton, and Searle (1973) reported, this strategy has serious consequences for test reliability and precision. Reaching an appropriate compromise between simple interaction and flexibility to review is, in large part, an empirical problem. This project employed conceptually simple techniques to allow students more flexibility for review.

A variety of techniques exist for computerized-adaptive testing. Empirical comparisions have been made between conventional paper-and-pencil tests and computerized-adaptive tests. Bryson (1971) found no difference between the two types of tests in predicting the score on the original longer test from which the items were derived. She suggested that the mode of item administration had an effect on the test score, although she was unable to identify that effect in her study. Bayroff and Seeley (1967), on the other hand, found the computerized-adaptive test to be more predictive of the longer test score. Hansen (1969) found both internal consistency and correlation with an achievement test criterion to be higher on a computerized-adaptive test than on a paper-and-pencil test. Two effects were present in these studies: (1) the effect of adaptive testing and (2) the effect of testing by computer rather than by paper-and-pencil.

Several studies have explored the effects of adaptive testing within a test administration medium. Three compared adaptive to conventional testing strategies using paper-and-pencil administrations (Bayroff, Thomas, & Anderson, 1960; Olivier, 1974; Wood, 1969). The results of these studies should be questioned. Although the conventional tests appeared to yield higher reliability and validity coefficients, low ability students were often omitted from the adaptive sample because they did not follow the instructions. When the conventional and adaptive tests were administered by computer, comparisons made by Betz and Weiss (1973, 1975), Larkin and Weiss (1974), and Vale and Weiss (1975) indicated conflicting results. The studies did not indicate uniformly higher reliability with one type of test.

A few studies have compared computer-administered to paper-and-pencil-administered tests. In a study of deaf students taking Raven's Progressive Matrices, Hitte, Riffer, and Stuckless (1971) found no significant effect on the mean test score that was attributable to administering the test in the traditional group presentation as opposed to a computer-managed presentation in which items were read from a booklet and responses were entered on a keyboard. Hitte et al. found the standard deviation to be slightly larger under computer-managed presentation.

Three studies, all using verbal tests, compared computer-administered to conventionally administered tests (Feurzeig & Jones, 1970; Serwer & Stolurow, 1970); Vinsonhaler, Molineux, & Rogers, 1968). All three studies found that scores were higher on the computer-administered test than on the conventionally administered test; the differences were greatest for low ability students. Fuerzeig and Jones also found the standard deviation to be greater on the paper-and-pencil test.

Objectives

Several questions were to be answered by this research:

1. What is the answer-changing behavior of students as they proceed through a computer-adminstered test?
2. Does the administration medium affect a) the estimate of the population mean? b) the population variance? c) the test reliability? d) the standard error of measurement?
3. Are the effects on the above test statistics dependent on the ability of the subject?
4. Is latency to respond related to the probability of a correct answer on the computer-administered version, and is this relationship a function of the quality of the item?

Method

Two experiments were performed in this study. In the first, subjects took parallel forms of a test administered by paper-and-pencil (PP) and by computer. In the second, subjects took tests of parallel forms administered by the same medium.

Instruments

Two tests were used in this study. One was the 50-item Vocabulary Test for High School Students and College Freshmen (Traxler, 1964), which requires the student to retrieve from memory the meaning of a given word and to select from among five alternatives the most appropriate definition. Subjects were given a time limit of 15 minutes to complete the test. The other test was the 30-item Symbolic Reasoning Test from the Employee Aptitude Survey (Ruch & Ford, 1957), which requires reasoning to determine the truth of a given conclusion. Subjects were given five minutes to complete this test. Two parallel forms of each test were used.

Administration Media

The paper-and-pencil versions were administered in the traditional manner; subjects were handed a two-sided test form and were to complete as many items as they could in the time alloted.

The computer-administered versions were intended to provide a similar approach to that used by most students on paper-and-pencil tests. The experiment was performed at the Naval Training Center, San Diego, using the PLATO IV computer system at the University of Illinois. The terminals used in this study were responsive to touch. Subjects were presented an item on the touch-sensitive plasma panel of the PLATO system and responded by touching a square on the panel. They could change their answers by touching a new square, go on to the next problem by pressing "NEXT," or return to the previous problem by pressing "BACK." Five sample problems were given. Specific off-line instructions were as follows:

> "When you take the tests on the computer, instead of blackening in an answer space with a pencil, you will simply touch an answer space on the screen with your fingertip. The terminal should respond

with a "beep," and an arrow will appear beside the answer you have chosen. If you miss the answer space by a little, an arrow will not appear, and you will have to touch it again. Back your fingertip away from the screen an inch or two after each response. If you want to change an answer, simply touch the square on the screen for the answer that you want. It is good procedure to make sure that an arrow appears beside the answer of your choice before going on. You can change a response as many times as you wish.

"One square on the screen will be marked NEXT and another marked BACK. To go on to the next item in a test, touch NEXT. To go back to any previous item, touch BACK. You may skip over a difficult item and return to it later if you wish.

"The Practice Sessions before each test are not timed and will help you to become familiar with these procedures. The Tests will be timed and will give you no clue as to the correct answer."

Students who completed the last problem within the time limit were told they had completed the test and were cycled through the test again, beginning with the first item. For each re-presented item, the last response (if one was made) was also presented. Students could press NEXT to continue through the second cycle with no alterations or they could touch a square to change their answer.

Experiment I

Design. A total of 118 recruits at the Navy Training Center, San Diego, participated in this experiment in the fall of 1975. Each subject was administered four tests: One form of each of the two tests described above was administered by computer, and a parallel form of each was administered using paper and pencil. The same time limits applied for both types of administration.

Figure 1
Experimental Design with Sample Sizes

	Vocabulary		Symbolic Reasoning	
First	First Form		First Form	
Administration	A	B	A	B
Computer	$16_{(1)}$	$15_{(3)}$	$17_{(5)}$	$12_{(7)}$
Paper and Pencil	$12_{(2)}$	$15_{(4)}$	$17_{(6)}$	$14_{(8)}$

Note. Design identifiers are in parentheses

The design was counterbalanced by test content, test form, and administration medium. The design and sample cell sizes are shown in Figure 1. Both computer administrations were taken successively, as were both paper-and-pencil administrations. Subjects who took the vocabulary test first on the computer also took the vocabulary test first using paper and pencil.

Groups 1, 4, 5, and 8 (Batch 1) took Form A on the computer and Groups 2, 3, 6, and 7 (Batch 2) took Form B on the computer.

Data were also available on the composite sum of three Basic Test Battery scores--Arithmetic Reasoning (ARI), General Classification Test (GCT), and Mechanical Aptitude (MECH). This composite sum is used by the Armed Services to classify individuals into mental categories. The percentiles of the mobilization sample (recruiting pool) included in each mental category are as follows:

Mental Category	Percentile
I	93-100
II	65-92
III	31-64
IV	10-30
V	0-9

Mental Category V applicants are accepted into the Armed Forces only during wartime. In this study an attempt was made to obtain equal numbers of subjects in each of the four mental categories I-IV. The distributions are shown in Table 1. There were missing data for a few students who were either called out of the testing session or who were being tested when the PLATO system failed.

Table 1
Distribution of Subjects by Mental Category for Experiment I

Mental Category	Batch 1 Form A:CA Form B:PP	Batch 2 Form A:PP Form B:CA
I	12	18
II	23	20
III	15	7
IV	11	12
Total	61	57
Mean Composite Sum	167	173

Experiment II

Design. Experiment II used the same tests as Experiment I with the same administration procedures. Instead of parallel forms being administered by different mediums, however, they were administered by the same medium. That is, a student took all four versions by computer or all four using paper and pencil. Presentations were counterbalanced for test content and test form.

A total of 91 subjects, similar to those in Experiment I, participated in Experiment II in the winter of 1976. The distribution of subjects by administration medium and mental category is given in Table 2.

Table 2
Distribution of Subjects by Mental Category
for Experiment II

Mental Category	Computer Administration	Paper-and-Pencil Administration
I	11	5
II	8	15
III	15	18
IV	9	10
Total	43	48
Mean Composite Sum	161	159

Results and Discussion

Flexible Computer-Administered Testing

Before reporting the test statistics resulting from the administration of a conventional paper-and-pencil test on the computer, it is necessary to discuss the extent to which the computer-administered procedure was successful in allowing students the same flexibility in changing answers as they would have on a paper-and-pencil test. This discussion includes all computer-administered tests from both experiments, separating results for the vocabulary and symbolic reasoning tests.

Table 3
Average Number of Answer Changes of Those Who
Changed Answers in Sequence

Mental Category	Experiment I		Experiment II	
	Form A	Form B	Form A	Form B
Vocabulary				
I	3.6	2.6	2.6	3.0
II	2.2	2.3	2.0	3.1
III	2.9	3.3	3.8	3.6
IV	3.4	2.9	3.3	3.2
Total	3.0	2.6	3.1	3.3
Symbolic Reasoning				
I	1.0	1.2	1.0	1.0
II	1.0	1.3	1.0	1.3
III	1.4	1.7	1.7	1.7
IV	2.3	1.3	1.5	1.0
Total	1.4	1.4	1.4	1.4

The vocabulary test was less speeded than the symbolic reasoning test. Over two-thirds of the students in any of the four test sets (two forms, two experiments) responded to at least one of the last three vocabulary items; whereas only 5% of the students in any test set responded to the last three symbolic reasoning items.

"In-sequence" changes occurred when the student changed his/her answer on the first presentation of an item before pressing "NEXT." "Out-of-sequence" changes occurred either when the student pressed "BACK" and was presented an item again or when the student completed the test and was automatically cycled through again, being re-presented each item until the time limit was reached. These changes are described by using the total number of answer changes, the proportion of subjects who changed at least one answer, and the average number of answer changes for those who changed at least one answer.

Answer changes during the sequence ("in sequence") were proportionately greater for the four sets of vocabulary tests (436 changes) than for the symbolic reasoning tests (106 changes). The vocabulary test averaged 3.0 answer changes for each subject who changed at least one answer (Table 3) compared to 1.4 on the symbolic reasoning test. This result reflects more than merely a longer test (50 vs. 30 items), since the proportion who changed answers on the vocabulary test (Table 4) was .76, compared to .39 on the symbolic reasoning test. The proportion of subjects who changed answers and the average number of answer changes for those who changed answers increased slightly as the ability level decreased for the symbolic reasoning test, but not for the vocabulary test.

Table 4
Proportion of Subjects Changing Answers in Sequence

Mental Category	Experiment I Form A	Experiment I Form B	Experiment II Form A	Experiment II Form B
Vocabulary				
I	.75	.71	.64	.64
II	.67	.70	.88	1.00
III	.80	.80	1.00	.87
IV	.64	.67	.78	.67
Total	.72	.70	.84	.79
Symbolic Reasoning				
I	.17	.28	.36	.27
II	.27	.45	.38	.38
III	.73	.43	.40	.47
IV	.36	.50	.44	.33
Total	.40	.40	.40	.37

The vocabulary test emphasizes memory retrieval, whereas the symbolic reasoning test emphasizes cognitive processing. On the vocabulary test, the subject who retrieves partial or no information and makes a selection could retrieve further information before pressing "NEXT." Thus, his/her answer-changing behavior would be increased on the basis of retrieval of new information. On the other hand, the symbolic reasoning test requires processing. Students who either realized the reasoning or recognized faulty reasoning before pressing "NEXT" could change their answers, but this was less likely to occur. An additional factor which may have caused this differential answer-changing behavior may be the number of response alternatives: five for the vocabulary test and only three for the symbolic reasoning test.

Out-of-sequence changes, in all probability made during the second cycle, show large differences. Because so few students completed the symbolic reasoning test, at most four students in a given set changed any answers out of sequence (Table 5) with a total of 13 answer changes across all sets.

Table 5
Average Number of Answer Changes of Those Who Changed Answers Out of Sequence

Mental	Experiment I		Experiment II	
Category	Form A	Form B	Form A	Form B
Vocabulary				
I	3.1	2.4	4.2	4.0
II	3.6	3.1	4.7	2.8
III	5.0	4.0	1.3	5.5
IV	2.0	2.4	1.3	---
Total	3.7	2.8	3.3	4.0
Symbolic Reasoning				
I	1.0	2.0	---	---
II	1.0	1.0	---	---
III	1.0	---	1.0	1.0
IV	---	1.0	---	2.0
Total	1.0	1.25	1.0	1.5

On the other hand, across all four sets, the vocabulary test resulted in an average of 3.4 out-of-sequence changes for those who changed at least one answer. This average was higher than the average number of in-sequence changes. Although there were no ability group differences for in-sequence vocabulary test changes, the proportion of subjects who changed answers out of sequence (Table 6) decreased from approximately .65 for the highest ability group to .26 for the lowest. But the average number of answer changes for those who changed at least one answer did not vary across ability levels.

Table 6
Proportion of Subjects Changing Answers Out of Sequence

Mental	Experiment I		Experiment II	
Category	Form A	Form B	Form A	Form B
Vocabulary				
I	.58	.59	.82	.64
II	.60	.70	.75	.63
III	.40	.40	.27	.27
IV	.18	.42	.44	.00
Total	.45	.57	.54	.37
Symbolic Reasoning				
I	.08	.06	.00	.00
II	.07	.10	.00	.00
III	.07	.00	.07	.07
IV	.00	.08	.00	.11
Total	.06	.07	.02	.05

Provisions for answer changing increase both the expense of computer-administered testing in terms of preparation and computer-processing requirements and the complexity of the man-computer interaction. A natural question is whether or not the increased cost and complexity are compensated by gains in test reliability and/or validity; it is thus relevant to investigate whether answer changes are random with respect to correctness or whether they increase or decrease the proportion of correct answers. The responses examined were those occurring at any time from the first response up to and including the last response. The relevant changes were right to wrong compared with wrong to right in all cases.

The changes by correctness of answer are shown for the two tests both in and out of sequence in Table 7. For all tests, there was a much larger proportion of in-sequence incorrect answers changed to correct than would be expected by chance. This indicates that students who decide to change their answers are not doing so randomly, but they do have some accurate knowledge that the changed answer is more likely to be correct. The out-of-sequence answer changes maintained this ratio of wrong-right to wrong-wrong changes. Furthermore, consistent with previous research (Mueller & Wasser, 1977), the number of wrong-right changes is consistently larger than right-wrong changes for both in-sequence and out-of-sequence changes. Thus, with the opportunity to review and change answers, students increased their scores. This increase was greater for the in-sequence changes, which may be partially due to accidental or impulsive touching of the panel. Generally, for the vocabulary test only, the upper ability groups were more likely to change wrong answers to right ones than were lower ability groups both in and out of sequence.

Table 7
Answer Changes by Correctness of Answer
Over Both Experiments and Forms

		In-Sequence		Out-of-Sequence	
	First	Last Answer		Last Answer	
	Answer	Right	Wrong	Right	Wrong
Vocabulary					
	Right	20	73	33	88
	Wrong	189	154	109	84
Symbolic Reasoning					
	Right	7	19	4	1
	Wrong	48	33	5	3

There was a suprisingly large number of right-right changes. Some of these may result from the student anticipating the next item after selection of an answer, thus repeating the answer before remembering to press "NEXT"; or the student may hesitate, plan to change the answer, and then decide upon the original selection again.

As students are cycling through the vocabulary test a second time, if they had high subjective confidence of the correctness of their answers,they are most likely to quickly confirm their response. If the students' confidence were lower, they might again attempt to retrieve from memory the necessary information. Accordingly, less confidence in correct answers appears in the out-of-sequence changes. On the vocabulary test there were more right-wrong out-of-sequence changes (88) than right-wrong in-sequence changes (73), but fewer wrong-right (109 out-of-sequence changes and 189 in-sequence changes).

However, as a student is cycling through the symbolic reasoning test a second time, he/she may reprocess the information. If the students' reasoning were faulty, he/she would be more likely to employ different reasoning processes on a second attempt. It is possible that an alternate reasoning method may result in the correct answer. Unfortunately, too few students cycled through the symbolic reasoning test a second time to obtain much information on answer changes out of sequence.

Given the opportunity to review and change answers, students utilized both in-sequence and out-of-sequence answer-changing features of a computer-administration system. To allow students to demonstrate performance similar to that on a paper-and-pencil test, a flexible system should be used. This flexibility appears to be more important for less speeded tests.

Effects on the Mean and Standard Deviation

Two test statistics which may be affected by presentation medium are the mean and the variance. Does the computer administration increase the mean and decrease the variance over a paper-and-pencil administration, as Feurzeig and Jones (1970), Serwer and Stolurow (1970), and Vinsonhaler, Molineux, and Rogers (1968) have reported?

Table 8

Means by Mental Category, Form, and Medium for Experiment I

Mental	Batch 1			Batch 2		
Category	Form A:CA	Form B:PP	*n*	Form B:CA	Form A:PP	*n*
Vocabulary						
I	37.5	38.1	12	42.1	41.9	16
II	33.7	36.3	23	35.0	34.4	20
III	23.4	25.2	15	24.4	26.8	5
IV	18.0	16.4	11	16.3	18.0	12
Total	29.1	30.4	61	31.9	32.2	53
Symbolic Reasoning						
I	14.9	15.7	12	14.1	16.6^a	17
II	14.5	15.8	22	11.6	14.2^b	20
III	8.1	11.1^b	15	10.4	11.1	7
IV	8.0	9.5	11	8.3	9.8	12
Total	11.8	13.4^b	60	11.5	13.6^a	56

$^a p<.05$

$^b p<.01$

The means by ability for Experiment I (shown in Table 8) are grouped by batches, with all students taking a given form-medium combination contributing to the mean. If the medium affects the mean, it would be expected that the difference in means for Batch 1 would vary in the same way as the difference in means for Batch 2. As indicated in the table, there was virtually no difference between means for either batch on the vocabulary test. On the symbolic reasoning test, however, scores were higher with paper-and-pencil administration across all ability levels regardless of test form; and the differences were statistically significant at several ability levels, despite the small number of subjects.

A comparison of means on the vocabulary test of the two forms in Experiment II (Table 9) indicates no differences on Forms A and B taken by the same individuals on the same medium and no difference on the same form taken by independent samples on two different mediums. Also, the symbolic reasoning test showed no difference in means on two different forms administered by the same medium. However, the symbolic reasoning test showed higher scores with paper and pencil in every case when the two samples took the test by two different media. In both experiments, overall scores on the symbolic reasoning test were over two points higher when the test was administered with pencil and paper. The difference was statistically significant, as Table 8 and 9 show, for all four comparisons (Form A with Form B in Table 8 and Form A with Form A and Form B with Form B in Table 9). These differences were consistent across ability levels.

Table 9

Means by Mental Category, Form, and Medium for Experiment II

Mental Category	Computer Form A	Computer Form B	*n*	Paper-and-Pencil Form A	Paper-and-Pencil Form B	*n*
Vocabulary						
I	40.8	41.3	11	37.8	39.8	5
II	34.5	33.5	8	36.0	35.1	15
III	22.3	21.5	15	21.8	22.2	18
IV	15.1	15.6	9	18.1	17.9	10
Total	27.8	27.6	43	27.1	27.2	48
Symbolic Reasoning						
I	14.7	14.5	11	19.0[a]	20.2[a]	5
II	10.5	11.1	8	14.8[b]	13.7	15
III	9.8	8.6	15	11.3	11.9	18
IV	6.0	5.3	9	9.0[a]	7.8	10
Total	10.4	9.9	43	12.7[a]	12.5[a]	48

[a] Mean PP = CA $p<.05$

[b] Mean PP = CA $p<.01$

These results suggest that administration medium on a vocabulary test, primarily requiring memory retrieval, does not affect the estimate of the population mean. On a symbolic reasoning test, which requires more analytical processing, there is a tendency for the computer-administered test to produce lower mean scores. Thus, the interaction of processing requirements and administration medium results in lower scores on a computer-administered test requiring processing and no difference in scores on a test requiring memory retrieval. An alternative possibility is that the administration medium only affects the estimate of the population mean on either speeded or difficult tests. The test-medium interactions are consistent across ability levels.

Table 10
Standard Deviations by Form and Medium for Experiment I

Test	Batch 1			Batch 2		
	Form A:CA	Form B:PP	*n*	Form B:CA	Form A:PP	*n*
Vocabulary	11.6	10.6	61	11.4	10.6	53
Symbolic Reasoning	4.4	4.7	60	4.3	5.1	56

The standard deviations for Experiment I shown in Table 10 and Experiment II shown in Table 11 indicate no difference between Forms A and B for the vocabulary test. However, across media, the paper-and-pencil test administration of the vocabulary test resulted in standard deviations almost one point lower than the computer-administered test for Experiment I (Form A compared with Form B and Form B compared with Form A in Table 10) and two points lower for Experiment II (Form A compared with Form A and Form B compared with Form B in Table 11). There were no consistent trends for the symbolic reasoning test.

Table 11
Standard Deviations by Form and Medium for Experiment II

Test	Computer			Paper-and-Pencil		
	Form A	Form B	*n*	Form A	Form B	*n*
Vocabulary	11.7	11.6	43	9.8	9.8	48
Symbolic Reasoning	5.2	4.9	43	4.7	5.5	48

The results of past studies that found lower ability students performing better on a computer-administered test were not supported here. Generally, the classification of ability in these studies is based on performance on one of the test versions. Given the score on the paper-and-pencil test, it would be expected that there would be a regression toward the mean. If, in addition, the mean were to be increased, the effect would appear greater for the lower ability students.

Classifying ability levels on the basis of an external criterion circumvents this problem. In the present case, however, it may have introduced an additional problem: The classification into mental categories is based on a composite of several dimensions. Although these dimensions are positively correlated, a better ability classification method appropriate for the given test would have used a verbal test for the criterion in assessing the vocabulary test effect and, perhaps, a logic test in assessing the symbolic reasoning effect.

It is important to note that the improvement for "lower ability" students is not for students of low general ability, but may be true only for students low in the ability being tested. If that is the case, the selection of administration medium on the basis of decreased test variance may result in less biased measurement. In either case, mental categories would not be used to interpret scores or to "correct" for the effects of administration medium.

Effects on Reliability

Reducing the standard deviation may reduce test reliability. The correlations between parallel forms presented by the two administration media from Experiment I are shown in Table 12, and parallel forms reliability coefficients within the same medium from Experiment II are in Table 13. Both tables also include the standard errors of measurement. The sample sizes for ability levels were too small to obtain accurate estimates and are, therefore, not presented.

Table 12
Correlations and Standard Errors by Form
and Medium for Experiment I

	Sample Size		Correlation		Standard Error	
Test	Group 1	Group 2	Group 1	Group 2	Group 1	Group 2
Vocabulary	61	53	.80	.93	6.5	4.1
Symbolic Reasoning	60	56	.60	.67	3.8	3.2

Note. Group 1 was administered Form A by computer and Form B by paper and pencil. Group 2 was administered Form B by computer and Form A by paper and pencil.

The reliabilities from Experiment II are approximately equal for the two media. This is true for both the vocabulary test and the symbolic reasoning test. Furthermore, for this population one of the tests had high reliability and the other had low reliability. Thus, the apparent lack of effect on reliability estimates attributable to computer administration may be consistent, regardless of the magnitude of the initial reliability estimate.

Table 13
Reliabilities and Standard Errors
by Test and Medium for Experiment II

	Sample Size		Reliability		Standard Error	
	CA	PP	CA	PP	CA	PP
Vocabulary	43	48	.91	.91	5.0	4.1
Symbolic Reasoning	43	48	.71	.74	3.5	3.7

Note. CA = computer administered; PP = paper and pencil.

Reliability, in this sense, represents consistency of measurement. Therefore, from Experiment II it can be said that abilities are measured as reliably, or consistently, by computer as by pencil and paper. However, there was an effect attributable to the administration medium; if there were no effect, the different medium correlations (Table 12) would closely approximate, if not equal, the reliability estimates from a single medium (Table 13). The correlations from Experiment I more closely approximate the reliability estimates from Experiment II on the vocabulary test than on the symbolic reasoning test. Thus, similar to the mean effects, there may be a processing by medium interaction. The more processing required, the more computer administration affects the behavior of the subject, although the effect is consistent from one form to the other. An alternative explanation, which is equally tenable, is that for less reliable tests, computer administration affects behavior more.

Latencies

Latencies were obtained for each student on each touch response of every item. If memory retrieval time or information processing time results in successful retrieval or processing, then a strong relationship between latencies and proportion correct would be expected. A log transformation was used on the latencies due to the nonlinearity of the measurement. The correlation between log latencies and the dichotomous variable, correctness (correct=1, incorrect=0) was computed for the first attempt on every item in Experiment II. The mean correlation of the 100 vocabulary items was -.35 and for the 50 symbolic reasoning items, with at least two respondents, was -.09. The large absolute mean correlation for the vocabulary test may indicate that if the necessary information is to be successfully retrieved during a reasonable amount of item test time, it should be retrieved quickly; if not retrieved quickly, it will likely not be retrieved at all. On the symbolic reasoning test, however, the longer the processing time, the more accurate will be the result.

The correlation between the log latency and correctness could be a function of the quality of the item, with the correlation being stronger for better items. The measure of quality used was the correlation between the item score and the total score. The relationship between the two correlations was measured by another correlation, which was found to be -.49 for the vocabulary test and -.34 for the symbolic reasoning test. Thus, for both tests the strongest relationship between latency and correctness occurred for better items and the weakest for poor items. Furthermore, this occurred more clearly for the vocabulary test than for the symbolic reasoning test, in all probability due to the weaker relationship between latency and correctness for the latter test.

Conclusions

This study has demonstrated some effects that may be found by administering a paper-and-pencil test on a computer system which allows the subject some of the flexibility offered by a traditional paper-and-pencil test. The effects depend upon the type of test administered and the subjects tested.

It was found that students utilized the opportunity to change answers: They did so more on the vocabulary test (which was less speeded and required memory retrieval) than on the symbolic reasoning test (which was very speeded and required information processing). Also, on the symbolic reasoning test the lower ability subjects who changed answers did so slightly more frequently than higher ability students. On the vocabulary test, however, during the first presentations, there were no differential answer changing rates by ability level. When items were reviewed out of the first presentation cycle, upper ability students were more likely to change answers on the vocabulary test than lower ability students.

While computer-administration of a paper-and-pencil test did not affect mean performance on the vocabulary test, it lowered performance on the symbolic reasoning test. However, the variance of the paper-and-pencil vocabulary test was slightly lower than that of the computer-administered version. No such effects were found for the symbolic reasoning test. The opposite effect on verbal tests, namely, lower variance on the computer-administered test, has been reported in other studies.

Reliabilities were found to be as strong for administration by computer as they were for administration by paper and pencil. Administration medium has an effect on test behavior, however; for the correlations across media were lower than within media. Latencies were found to be related to correctness more for the vocabulary test than for the symbolic reasoning test, and the relationship was shown to be stronger for higher quality items.

Generally, it was found that students adequately utilized the option of answer changing and review on a computer-administered version of a paper-and-pencil test. Furthermore, test statistics were only minimally affected by computer administration. Thus, the vast capabilities that are provided by the use of computers in adapting tests to students' abilities and in providing feedback can be used with minimal interference from the computer medium presenting items. This does not imply that such a feature is worth having. It is possible that more accurate ability estimates can be made by administering new items rather than reviewing the old ones.

Further research efforts should be designed to determine whether the effects of administering a paper-and-pencil test by computer are dependent on

1. The amount of processing required,
2. The speededness of the test,
3. The reliability of the initial test,
4. The difficulty level of the initial test, and
5. The number of response alternatives.

References

Bayroff, A. G., & Seeley, L. C. An exploratory study of branching tests (Technical Research Note 188). U. S. Army Behavioral Science Research Laboratory, June 1967.

Bayroff, A. G., Thomas, J. J., & Anderson, A. A. Construction of an experimental sequential item test (Research Memorandum 60-1). Department of the Army, Personnel Research Branch, January 1960.

Betz, N. E., & Weiss, D. J. An empirical study of computer-administered two-stage ability testing (Research Report 73-4). University of Minnesota, Department of Psychology, Psychometric Methods Program, 1973.

Betz, N. E., & Weiss, D. J. Empirical and simulation studies of flexilevel ability tests (Research Report 75-3). University of Minnesota, Department of Psychology, Psychometric Methods Program, 1975.

Bryson, R. A comparison of four methods of selecting items for computer-assisted testing (Technical Bulletin STB 72-8). San Diego, CA: Naval Personnel and Training Research Laboratory, 1971.

Feurzeig, W., & Jones, G. Reevaluating low achievers with computer-administered tests. Unpublished report. Bolt Beranek & Newman, Inc., 1970.

Hansen, D. N. An investigation of computer-based science testing. In R. C. Atkinson & H. A. Wilson (Eds.), Computer-assisted instruction: A book of readings. New York: Academic Press, 1969.

Hansen, D. N., Johnson, B. F., Fagan, R. L., Tam, R., & Dick, W. Computer-based adaptive testing models for Air Force technical training. (Final report for contract F41609-73-C-0013). Lowry Air Force Base, CO: Air Force Human Resources Laboratory, December 1973.

Hitti, F. J., Riffer, R. L., & Stuckless, E. R. Computer-managed testing: A feasibility with deaf students. National Technical Institute for the Deaf, July 1971.

Jensema, C. J. An application of latent trait mental test theory to the Washington pre-college testing battery. Unpublished doctoral dissertation, University of Washington, 1972.

Larkin, K. C., & Weiss, D. J. An empirical investigation of computer-administered pyramidal ability testing (Research Report 74-3). University of Minnesota, Department of Psychology, Psychometric Methods Program, 1974.

Lord, F. M. Some test theory for tailored testing. In W. H. Holtzman (Ed.), Computer-assisted instruction, testing, and guidance. New York: Harper & Row, 1970.

Mueller, D. J., & Wasser, V. Implications of changing answers on objective test items. Journal of Educational Measurement, 1977, 14, 9-14.

Oliver, P. An evaluation of the self-scoring flexilevel tailored testing model. Unpublished doctoral dissertation, Florida State University, 1974.

Ruch, F. L., & Ford, J. S. Employee Aptitude Survey. Los Angeles: Psychological Services, Inc., 1957.

Serwer, B. L., & Stolurow, L. M. Computer-assisted learning in language arts. *Elementary English*, 1970, *47*, 641-650.

Suppes, P., Fletcher, J. D., Zannotti, M., Lorton, P. V., & Searle, B. W. *Evaluation of computer-assisted instruction in elementary mathematics for hearing-impaired students* (Technical Report 200). Stanford, CA: Stanford University, Institute for Mathematical Studies in the Social Sciences, March 1973.

Traxler, A. E. *Vocabulary Test for High School Students and College Freshmen*. Indianapolis, IN: Bobbs-Merrill Co., Inc., 1964.

Vale, C. D., & Weiss, D. J. *A study of computer-administered stradaptive ability testing* (Research Report 75-4). University of Minnesota, Department of Psychology, Psychometric Methods Program, 1975.

Vinsonhaler, J. F., Molineux, J. E., & Rogers, B. G. An experimental study of computer-aided testing. In H. H. Harman, C. E. Helm, & D. E. Loye (Eds.), *Computer-Assisted Testing Conference Proceedings, November, 1966*. Educational Testing Service, Princeton, NJ, 1968.

Weiss, D. J., & Betz, N. E. *Ability measurement: Conventional or adaptive?* (Research Report 73-1). University of Minnesota, Department of Psychology, Psychometric Methods Program, 1973.

Wood, R. The efficacy of tailored testing. *Educational Research*, 1969, *11*, 219-222.

DISCUSSION: SESSION 8

DAVID J. WEISS
UNIVERSITY OF MINNESOTA

The study by Sachar and Fletcher, while it is interesting, is a good example of the fact that the use of large time-shared computer systems for adaptive testing is likely to be detrimental to the future of adaptive testing. Sachar and Fletcher were interested in the question of whether or not it is appropriate to administer standard paper-and-pencil tests by a computer. This is a good question; however, I do not think the answer given by Sachar and Fletcher is a good answer.

In the first test, a verbal reasoning test, there were 50 items with a 15-minute time limit; this was quite reasonable, since many of the people reached the last question. It was an untimed power test. In the second test, which was a symbolic reasoning test, there were 30 items with (although it was not indicated) a time limit of 5 minutes, or 10 seconds per item. In any large-scale time-shared system that I have seen, system response time has been unpredictable. Consequently, from the 10 seconds maximum per item can be subtracted an unknown response time which, on the average, could run from a half-second to 2, 3, or 5 seconds. The time limit for each item was thus reduced from 10 seconds to about 8 seconds. The testee's reaction time must also be considered. These timing factors could explain the result that the second test was more difficult than the first test. Using number correct scores (which is a non-optimal way to measure ability) on a heavily timed test is inappropriate, since testees cannot answer as many items in a given unit of time (say, 5 minutes) with a computer display as they can with paper-and-pencil. Consequently, there will be a mean difference in number correct scores, which is what Sachar and Fletcher found.

Why did the low ability people have a mean difference that was greater than the high ability people? It is known that there is a relationship between response latencies and ability and that there is a relationship between correctness of responses and latencies. If the low ability people (who tend to work slower) had had the opportunity, perhaps they could have completed all the items. However, they answered fewer items on the computer because there were fewer items to answer in that period of time. It is therefore likely that the results, rather than being a function of cognitive information processing, are a function simply of the time limits and the system response time that was not accounted for in the study.

With regard to the answer-changing data, it is neither particularly impressive nor particularly interesting. Research in computerized adaptive testing should not even be concerned with answer-changing behavior. Answer-changing behavior is a characteristic of the paper-and-pencil testing context. People change answers because they are able to change answers. The way to investigate answer-changing behavior is to look at things such as whether or not the answer changes lead to higher reliability, higher validity, or greater information in the scores, not the number of changes or the direction of the changes. The fact that people change answers is irrelevant in computer-administered testing. In adaptive testing the testee cannot really be given the opportunity to change answers. The only reason I can see for allowing people to change answers in computer-administered tests is that of public relations. Whether or not people feel better because they have been given the opportunity to change answers is an empirical question. We have administered approximately 10,000 adaptive tests, and I do not think more than one-half of 1% of the people have complained about not being able to change their answers.

To carry Sachar and Fletcher's research to its logical extreme, how many times do we let testees go through a test? They can go through a paper-and-pencil test 2, 3, 5, 10, or 20 times and change their answers until the eraser falls off their pencil. But if they were allowed to go through the test only once, it might make them more unhappy than if they were allowed to go through it 10 times, or not at all. I do not see any real conclusions to be drawn from this paper other than that the second test was too difficult and too speeded and that Sachar and Fletcher found what they expected to find, because they did not take into account other factors that might influence response time.

With regard to Hunter's paper, I am happy to hear that it was possible to expand the computer system on the basis of the preliminary research, and I hope it will be possible to expand it further. However, I would suggest not buying a big computer system. I am disturbed by some of the things that are implicit in the use of large-scale computer systems for adaptive testing. In implementing computerized testing, whether it is adaptive or non-adaptive, many of us are too worried about the hardware requirements. Implementing computerized testing requires minimal computer hardware, and I think we run the risk of demanding too much hardware for its implementation. I just do not think large amounts of hardware are needed unless they will be used for some other purpose. At the same time, I do not think a good job of computerized testing can be done if the processors are being used simultaneously for other purposes. Thus, adaptive testing should use small processors dedicated to testing so that other applications do not interfere. Brian Waters' move toward the micro-terminal is a step in the right direction. Fixating on large processors requires too much money, too fast, when all that hardware is not really needed.

I have one comment about the factor analysis that Hunter reported, in which he was trying to show that the computer-administered tests measured something different than the paper-and-pencil battery. To some extent Hunter has overdrawn his conclusions because some of the variables that he used in his factor analysis were scores such as the number of errors in Minute 1, the number of errors in Minute 2, the number of errors in Minute 3, the number of errors in Minute 4, and number of errors in Minute 5, all on the same test. Obviously, if those variables are intercorrelated, they will correlate so highly among themselves that they must define a factor which is unique to the computer-administered test. To be fair to the question being asked, if one wants to do those kinds of factor analyses, the factor results should not be overdetermined by putting in highly redundant scores--scores derived from the same test and, in all likelihood, correlated very highly with each other, so that it would be almost impossible for them to load on any other factor.

With regard to Hunter's scoring, I think that some of the scoring approaches are too traditional, too simplistic, and lose too much information. The number of misses that a person makes in a certain unit of time is a variable which is easy to count. But now that the capability of using the computer exists and, consequently, the possibility of obtaining more information, recording more information, and utilizing it in a better way, why use simple scores such as the number of misses in a given unit of time when much more information can be extracted? Some possible scores that could be used are the distance that the individual was from the target value, how long the person was at that distance, and how long it took him/her to get back to the target value as a function of the amount of time that the target value was there. When those kinds of scores are combined, either implicitly or explicitly, with some of the models discussed earlier (e.g., the continuous latent trait model, in which scores of that nature which are inherently continuous can be adequately utilized), they will provide much more information with which to adequately demonstrate that computerized testing, whether it is adaptive or non-adaptive, will contribute to prediction. Meanwhile, much time is being spent improving tests and poor criteria are being used. For example, a pass/fail criterion can be a gross and unreliable indicator which is very difficult to predict.

In order to have computerized testing accepted, as much time as possible must be spent refining criteria to make them relevant to the constructs measured by the predictors and the tests. We need to have parallel groups working in this area: one working on the predictor set developing tests and another refining criteria. Hunter's correlations of .20 and .30 with the dichotomous criterion are not high enough to impress future users of adaptive testing. The immediate goals of Hunter's research are fine, demonstrating that computerized testing is contributing something without being adaptive. However, thought should be given to ways of making these perceptual-motor tests adaptive. One way of doing that for the simple two-hand coordination test is by a real-time monitoring of what the person is doing in his/her interaction with the moving target. The distance that the person is from the target could be monitored on a continuous basis, and the latencies of the length of time it takes the individual to adjust to changes in the target could be computed. Then, on the basis of a continuous real-time analysis of that data as it is obtained, the test could be adapted to the individual. The target would then be moved faster as the person gets closer to the target. If he/she is the kind of person who appears to have good reaction time, the target could be moved away a little more quickly than it would be for someone who appears not to have a good reaction time--until the person's capabilities are ascertained. I do not think this will emerge from the kinds of models with which most of us are working. I think we will have to become more creative and see precisely how real adaptive, highly dynamic kinds of tests can be developed that will predict to real-life criteria.

PANEL DISCUSSION:

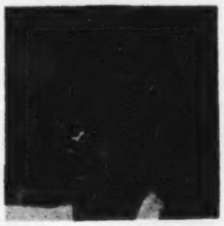
FUTURE DIRECTIONS FOR COMPUTERIZED ADAPTIVE TESTING

CHAIRMAN: MARSHALL J. FARR, OFFICE OF NAVAL RESEARCH

FREDERIC M. LORD
EDUCATIONAL TESTING SERVICE

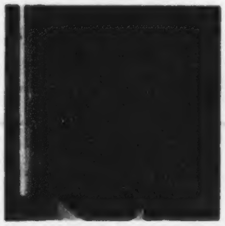
MARK D. RECKASE
UNIVERSITY OF MISSOURI-COLUMBIA

FUMIKO SAMEJIMA
UNIVERSITY OF TENNESSEE



VERN W. URRY
U.S. CIVIL SERVICE COMMISSION

DAVID J. WEISS
UNIVERSITY OF MINNESOTA



INTRODUCTION

MARSHALL J. FARR
OFFICE OF NAVAL RESEARCH

Since the first conference on computerized adaptive testing (CAT), which was held in Washington, D.C., slightly over two years ago and co-sponsored by ONR and the U.S. Civil Service Commission, there has been a great increase in the number of contractors working directly in the area of CAT not only for ONR but for the other military services as well. Five years ago, ONR began supporting the CAT research of David Weiss and his University of Minnesota group. The ONR program has expanded so that we now also support such researchers as Norman Cliff (University of Southern California), Mark Reckase (University of Missouri--Columbia), and Fumiko Samejima (University of Tennessee); and if our research resources permit, there will be some modest expansion. Other Navy and military activities, namely the Navy Personnel Research and Development Center (NPRDC), the Army Research Institute for the Behavioral and Social Sciences (ARI), and the Technical Training Division and Personnel Research Division of the Air Force Human Resources Laboratory, have embarked on in-house or contract-support research and development in this area.

In addition, two very recent Ph.D.'s who had received ONR support at the University of Minnesota have left the university world to spread the "CAT gospel" elsewhere. David Vale's new job will involve him with the testing enterprise of the State of Minnesota, and James McBride has become a civilian researcher for the ARI.

Since the first CAT conference, ONR and ARPA have started to sponsor work on computerized adaptive *achievement* testing. David Weiss' group extended their research from the ability testing domain to achievement testing; NPRDC and the other Services have now joined with ONR and ARPA in funding the University of Minnesota contract on adaptive achievement testing. The current ONR-sponsored work by Mark Reckase also focuses on this area.

There also exists an informal joint military service committee which will attempt to coordinate (and possibly collaborate on) the application of interactive computer technology to the total personnel-accession system in the military. That system includes not only the functions of recruiting, selection and classification testing, and assignment but also vocational guidance and counseling. A subcommittee of this joint service group has been set up to concern itself solely with adaptive testing. It has already met and will continue to do so on a regular basis to exchange information and coordinate plans and programs.

Finally, it should be noted that the research program at ONR is much broader ranging than CAT and psychometrics. For example, there are multi-contract programs in areas such as generative, knowledge-based computer-assisted instruction and individual differences in human information processing. Annual meetings of contractors are held in different thematic areas; they are informal get-togethers at which each contractor describes his or her latest research findings and plans. The only "outsiders" invited are a handful of scientists representing other appropriate Navy, Department of Defense (DoD), and federal activities. The program in adaptive testing is now large enough to warrant holding such a meeting for this thematic area. Through attendance at these meetings, military and civilian government scientists will gain first-hand information about the basic research we are supporting in CAT. Thus the DoD efforts in this area, especially the ONR-sponsored ones, are both widely coordinated and integrated.

FREDERIC M. LORD
EDUCATIONAL TESTING SERVICE

A few years ago, a well-known person told me about a study at Harvard which showed that if the best items from the Scholastic Aptitude Test (SAT) were selected and if proper scoring weights were used for these, the result would be a test as valid as the regular SAT and only half as long. I was impressed with this and said as much. His reply was, "Hell, who cares?" This is not the attitude of the Civil Service Commission, it is not the attitude of the military, but it is the attitude of some organizations that administer tests.

About a year ago, I made a resolution that I would not make predictions about the future of tailored testing. I intend to keep this resolution. I should like to discuss several points, however, which I think are important. First, prior to this conference, the problem of multidimensionality seemed to be the biggest problem in tailored testing. The major limitation of tailored tests at present is that the unidimensionality assumption restricts their use to a relatively small number of tests. Sympson's paper, however, indicates that a point may now have been reached where something practical can be done about this. For example, his multidimensional model can perhaps be used with a chemistry achievement test that contains chemistry information items and arithmetic reasoning items. Predictions about this kind of test from Sympson's model could be compared with predictions from a unidimensional model; the predictions from a multidimensional model should be better.

On the other hand, the SAT, which contains at least three different kinds of verbal items, could be analyzed by both a unidimensional and a multidimensional model. The unidimensional model might do almost as well as the multidimensional model. Conceivably, it could do better; when a large number of parameters are estimated, many degrees of freedom are lost and there is more sampling error.

Much has been said recently about examinee rights. People who take tests are interested in preserving their rights in the situation. What can be said to a minority examinee who has taken a test for a government position and questions the final hiring decision, saying that he/she answered more questions correctly on the tailored test than the person who was hired? If he/she were told that the person who was hired took items that were more difficult, he/she might respond by saying, "If you had given me the more difficult items, I would have done better than the person who was hired." Are we going to be able to convince the minority examinee and his/her lawyer and the judge that this is really not the case? Perhaps there will be a tendency to consider latent trait theory and item characteristic curve theory to be a way of complicating things so that the average person who has to take tests cannot understand what is going on and thereby suffers.

The final point I should like to make concerns validity studies in which test scores are correlated with grade-point average or some other external criterion rather than true score. Suppose a tailored testing procedure raises test reliability from .90 to .92. A typical validity coefficient is probably not much higher than .60; it could be quite a bit lower. Let us substitute these numbers into the classical formula showing the effect of changes in test reliability on the validity coefficient. How much will the gain in reliability from .90 to .92 raise the validity? When the conventional test has a validity of .60, the tailored test will have a validity of .6066.

This increment of validity from .60 to .6066 must be important. The conventional test would have to be lengthened by more than 25% to achieve such a gain. If we did not think such gains were important, we would be building shorter tests than we do.

There are numerous requests for validity studies to determine whether or not the tailored test is better than the conventional test, and if so, how much better. Suppose a validity study with 100 cases is performed. The standard error of a correlation coefficient of .60 is then .064. If the two tests being compared are given to two different groups, the standard error of the difference between two such correlation coefficients is about .09. Therefore, such a validity study is obviously a waste of time if the validity difference one is attempting to discover is the difference between .60 and .6066. If 1,000 cases are used in each group, this cuts the standard error by about one-third, so the standard error would then be about .03. If 10,000 cases are used, this cuts the standard error by one-third again. Now the standard error of the difference between the two validity coefficients is about .01; but this still is not good enough to detect a difference of .0066. There is still one chance in four that the poorer test will show up in the sample as better and the better test will show up as poorer.

The situation is much improved if the same examinees are given both tests under comparison. The extent of the improvement depends on the correlation between the two tests. Formulas are available for the large-sample standard error of the difference between two correlated correlations (validity coefficients). Validity studies should not be undertaken or reported unless such formulas show that the sample size used is large enough to ensure that the better test will be able to demonstrate its superiority in spite of sampling fluctuations.

MARK RECKASE
UNIVERSITY OF MISSOURI--COLUMBIA

The question "Who cares about latent trait theory and tailored testing?" has been raised earlier in this conference. In my experience, two large populations--teachers and pupils--do care about these fields but without knowing it. In general, negative feelings are expressed toward measurement because these groups have found it necessary to use techniques that are not seen as congruent with their goals. These goals currently are to humanize education while striving to right what are considered to be earlier wrongs perpetuated by measurement policies. The invasion of privacy issue as related to measurement is an expression of this trend. However, the new technology being discussed at this conference can bring about a more positive feeling by making testing more compatible with present educational goals. In this sense, students, teachers, and all others concerned with educational programs would care about the possibilities opened to them by these procedures, if only they were aware of their existence.

How can tailored testing be used in conjunction with educational programs to make them work better? One direction has been suggested earlier by Ernst Rothkopf when he stated that rather than looking at one little unit, the global entity should be perceived. Tailored testing is now one of those little units that may in the future be a component of a global computerized educational system. I can foresee that 20 years in the future there will be a marriage between computer-assisted instruction (CAI), computerized adaptive testing (CAT), and computer-managed instruction (CMI) to form this global system.

Tailored tests would fit beautifully into such a system. They are short, they are quickly administered, and they do not have to be scheduled. Numerous short exams can be built into the instructional process. Educational psychologists have been telling us for a long time that frequent testing with immediate feedback is optimal for efficient learning. Procedures are finally available to make this possible.

A second possible direction for the development of tailored tests in educational programs is to improve the prediction of performance. Current predictive procedures based on aptitude tests are, to a certain extent, a waste of time. The results are so general that teachers have a hard time interpreting the results. They ask, What do I do with an aptitude test? What is an intelligence test for? It is difficult to give a really good answer unless there is a specific criterion to be predicted, and even then, the predictive validity is probably only about .5.

Perhaps better predictors of future performance on various tasks can be obtained by using information about how a person performed in past tasks. In other words, how persons learned in past situations will predict how they will perform in future situations. Tailored testing can provide much data concerning this, especially if combined with CAI programs. For example, many short tailored tests can be used in educational programs, such as training in the Navy, Army, and Air Force. From these tests, growth curves can be obtained and the speed of learning computed for an individual. This may be an important variable for future predictions.

If these tests are self-paced, motivational components can be obtained as another predictive variable. How often testees are willing to take a test can be ascertained. If testees put off all of the tests to the end of instruction, they are not very highly motivated. If they come in to start taking tests immediately, they are extremely motivated. Motivational information in this operationally defined sense can thus be obtained. From these situations can be obtained more realistic measures of performance than are currently available from traditional aptitude tests, possibly improving predictions.

Based on the new technology being described at this conference, the classrooms of the future can be envisioned with good inexpensive cathode ray terminals (CRTs) built into every desk. Whenever a person feels he/she knows the material, he/she can just answer a few questions on the CRT, and a decision procedure will operate indicating whether or not the knowledge desired in the area has been acquired in order to progress to the next area; the hierarchical schemes discussed by Ronald Hambleton--the criterion-referenced models--will be part of this system. Perhaps 20 or 30 years from now, should everything continue to move along at the same pace, there will be a global model in existence that will revolutionize the educational system.

FUMIKO SAMEJIMA
UNIVERSITY OF TENNESSEE

There are at present a number of general tendencies in computerized adaptive testing research which can be discussed in terms of the problems they present and recommendations for their resolution.

Use of the Correlation Coefficient

It seems to be quite common that psychologists like to use the product-moment correlation coefficient in their research, even to the extent that it violates its logical limitations. It should be kept in mind that, in spite of its simplicity, the correlation coefficient is often misleading.

Suppose, for example, there are two methods of estimating the difficulty parameter of the test item, and the relative superiorities of the methods with respect to the accuracy of estimation are to be decided, using a set of available data. It appears to be a fairly common procedure for researchers to calculate two correlation coefficients: One is between the set of true difficulty parameters and the set of their estimates obtained by the first method, and the other is the counterpart using the estimates obtained by the second method. Then, it is decided that whichever has produced a higher correlation coefficient is the better method, provided that it is a monte carlo study and the true parameter values are known.

Let it be assumed that there are four test items; and their difficulty parameters, b_1, b_2, b_3, and b_4, are 2.0, 5.0, 3.0, and 0.0, respectively. Let it further be assumed that by the first method, -6.0, 0.0, -4.0, and -10.0, respectively, were obtained as their estimates; and by the second method, 0.0, 5.0, 3.0, and 2.0. It can easily be seen that the second method produced the exact parameter values for Items 2 and 3 and that, as a whole, these estimates are closer to the true parameter values than those obtained by the first method. If the two correlation coefficients are computed, however, the correlation for the second method is .692, while that for the first method is 1.000 (in other words, a perfect correlation). Thus, it is clear that the correlation coefficient can by no means be considered as an indicator of the accuracy of estimation. In fact, the perfect correlation only indicates that the configuration of the estimates is exactly the same as the configuration of the true parameters. Although this is a necessity for perfect estimation, it is far from sufficient.

To replace the correlation coefficient, the mean square error or its square root can be used in the above example, defining the error as the discrepancy between the estimate and its true parameter value. The mean square error and its square root for the first method are 59.5 and 7.714, respectively; these values are far greater than the respective values of 2.0 and 1.414 for the second method. Thus, the second method is far better than the first method in the accuracy of estimation.

As the above example shows, correlations can be misleading if they are used blindly. It is strongly suggested that researchers should stop and consider whether or not the use of a correlation coefficient is justifiable in the particular process of research before it is actually used.

Use of the Test Information as the Criterion for Terminating the Presentation of New Test Items

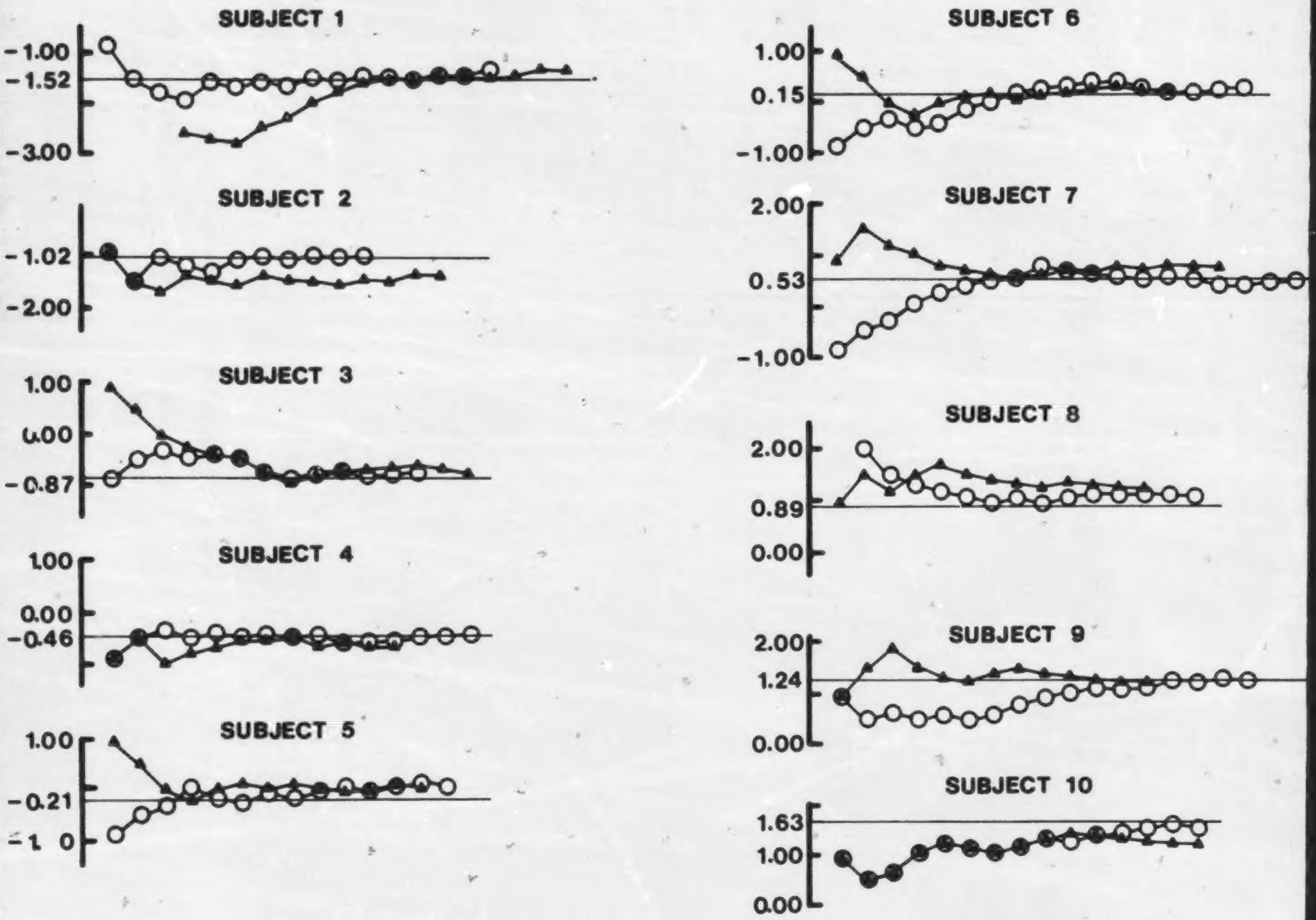
It seems to be common for researchers to apply the convergence of the current maximum likelihood estimate obtained after presenting each test item sequentially as the criterion for terminating the presentation of new items in computerized adaptive testing. This procedure, however, will result in producing different levels of accuracy of estimation at different levels of ability, or even at the same level of ability.

To illustrate this point, Figure 1 presents 10 graphs, each of which displays the process of convergence of the maximum likelihood estimate in a simulated tailored testing situation in each of two sessions (Samejima, 1977a). The ability level of each of these 10 hypothetical examinees is shown by a number on the ordinate and the attached horizontal line. For each examinee in each session, binary items were selected and presented until the test information at the current value of the maximum likelihood estimate had reached 25. Since the items are binary, no local maximum likelihood estimate was obtained after administration of the first item. For Subject 1, for instance, in the first session the first local maximum likelihood estimate was given after administration of the second item; and in the second session it was obtained after administration of the fifth item. It is clear from this figure that in some cases, the current maximum likelihood estimates converged well before the test information reached 25.0; whereas in other cases, they had not converged yet by the time the test information reached 25.0.

Consider, for example, Subject 4 in the first session (hollow circles) and Subject 10 in the second session (solid triangles). If the rule is made that the presentation of new items is to be terminated when the shift of the current maximum likelihood estimate is less than .07 twice in succession, then this will occur after the presentation of the 9th item in the former and not until the presentation of the 14th item in the latter. The corresponding values of test information are 13.370 and 23.137, respectively. The standard error of estimation, which is the inverse of the square root of test information, is .273 in the former and .208 in the latter, i.e., approximately 76% of .273 . On the other hand, if the rule is made that the presentation of new items is to be terminated when test information has reached, say, 25.0 , at that current maximum likelihood estimate, as was the case here, the standard error of estimation would be approximately the same for all the examinees of different ability levels, i.e., .20. If estimation of each examinee's ability with the same level of accuracy is desired, there will be no doubt that the second rule is better than the first rule.

If the same level of accuracy of estimation is unnecessary, as in selection, it will be possible to pre-arrange a desirable test information function which is not constant for the entire range of ability in question but has a specific curve for the specific purpose. This test information

Figure 1
Graphic Presentation of the Change of the Local Maximum Likelihood Estimate After the Presentation of Each New Item for Each of 10 Subjects



function can then be used as the criterion for terminating the presentation of new items. In such a case, examinees of different levels of ability are measured with different levels of accuracy of estimation and yet the resulting selection will be conducted as accurately as is desirable if the appropriate information curve is used.

The above are only two examples of many possibilities. In any case, the use of test information functions as the criterion for terminating the presentation of new items in tailored testing permits control of the level of accuracy of estimation to serve the purposes of testing; it is impossible to do so if the convergence of the current maximum likelihood estimate is used as the criterion. Therefore, the adoption of the test information function as the criterion is strongly recommended, rejecting the convergence of the current maximum likelihood estimate, which makes the accuracy of estimation arbitrary.

Elimination of the Reliability Coefficient and Use of the Standard Error of Estimation

It was unfortunate that classical test theorists defined the reliability coefficient of a test as the correlation coefficient between the two sets of test scores obtained by the repeated administrations of the same test or the administration of two parallel tests. In so doing, the impression was given that the reliability coefficient is an intrinsic property of the test itself. If the nature of the correlation coefficient is considered, it is evident that it depends heavily upon the particular group of examinees taking the test, as well as upon the test itself. I have called the reliability coefficient a dead concept in one of my published papers (Samejima, 1977a), pointing out the fact that not only is it a misleading concept but also a highly unnecessary one.

Especially in computerized adaptive testing, there is no need to use the test score (which is defined either as a weighted or an unweighted sum of item scores), since the maximum likelihood estimate (or some other estimate, statistically far more sophisticated than the test score) is readily obtainable from the original response pattern. In such a setting, the reliability coefficient will be replaced by the correlation coefficient between the two sets of maximum likelihood estimates, instead of the test scores, which are obtained by the repeated measurements.

Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be the maximum likelihood estimates of ability θ obtained by the repeated measurements, respectively. It has been shown (Samejima, 1977a) that such a correlation coefficient, $\rho_{\hat{\theta}_1\hat{\theta}_2}$, is perfectly predictable from the test information functions and the variance of the maximum likelihood estimate obtained by first administration, provided that the test information function assumes high enough values throughout the range of ability in which the total group of examinees are located so that the conditional distribution of the maximum likelihood estimate, given ability, approximately follows $N(\theta, I(\theta)^{-1})$. In computerized adaptive testing, this will be realized if some pre-arranged test information function (which is high enough throughout the range of ability in question) is used as the criterion for terminating the presentation of new items. Under such circumstances,

$$\rho_{\theta_1\theta_2} = [\mathrm{Var.}\ (\hat{\theta}_1) - E\{I^{(1)}\ (\theta)^{-1}\}] \qquad [1]$$

$$[\mathrm{Var.}\ (\hat{\theta}_1)\ \{\mathrm{Var.}\ (\hat{\theta}_1) - E[I^{(1)}(\theta)^{-1}] + E[I^{(2)}(\theta)^{-1}]\}]^{-\frac{1}{2}}\ ,$$

where $I^{(1)}(\theta)$ and $I^{(2)}(\theta)$ are the two criterion test information functions in the repeated measurements.

From Equation 1, it is obvious that this coefficient is at the mercy of the variance of the maximum likelihood estimate of ability for the particular group of examinees who happened to take the test. As long as this coefficient is accepted as the reliability coefficient of the test, however, it will be easy to deceive the public by making a badly constructed test look good, if the value of the reliability coefficient is increased by choosing a highly heterogeneous group of people as the examinees. Following a similar logic, it will also be easy to make a well-constructed test look bad if a group of examinees is chosen who are close to each other with respect to ability tested.

It should be noted that Equation 1 includes only one variance of the maximum likelihood estimate, which implies that, in practice, only one administration of the test is needed to compute this coefficient. In a special case in which the same test information function is used as the criterion in the two adminstrations, Equation 1 is simplified to the form

$$\rho_{\hat{\theta}_1\hat{\theta}_2} = [\mathrm{Var.}\ (\hat{\theta}_1) - E\{I(\theta)^{-1}\}]\ [\mathrm{Var.}\ (\hat{\theta}_1)]^{-1}\ , \qquad [2]$$

where $I(\theta)$ is the common criterion test information function. In practice, in both Equations 1 and 2 the expectation of the inverse of each test information function must be estimated in some appropriate way. When the criterion test information function is set to be a constant in each administration (i.e., the presentation of new items is terminated when the test information function has reached a certain constant value, regardless of the value of the current maximum likelihood estimate), Equations 1 and 2 are further reduced, respectively, to

$$\rho_{\hat{\theta}_1\hat{\theta}_2} = [\mathrm{Var.}\ (\hat{\theta}_1) - \sigma_1^2]\ [\mathrm{Var.}\ (\hat{\theta}_1)\{\mathrm{Var.}\ (\hat{\theta}_1) - \sigma_1^2 + \sigma_2^2\}]^{-\frac{1}{2}} \qquad [3]$$

and

$$\rho_{\hat{\theta}_1\hat{\theta}_2} = [\mathrm{Var.}\ (\hat{\theta}_1) - \sigma^2]\ [\mathrm{Var.}\ (\hat{\theta}_1)]^{-1}\ , \qquad [4]$$

where σ_1 and σ_2 in Equation 3 are the square roots of the inverses of the two constant information functions and

σ in Equation 4 is the square root of the inverse of the common constant test information.

Thus, from these four equations, it is obvious that the so-called reliability coefficient depends upon both the test information functions (or function) and the variance of the maximum likelihood estimate for the particular group of people who happened to be selected as the group of examinees. For this reason, it is advisable to discard the reliability coefficient and use the standard error of estimation instead.

It should be noted, however, that the standard error of measurement defined in classical test theory is no better than the reliability coefficient and that, in fact, it is closely related to the reliability coefficient. It is the standard error of estimation defined in latent trait theory which should replace the reliability coefficient. This standard error of estimation is defined as the square root of the inverse of the test information function; it does not depend upon any specific group of examinees but is solely dependent upon the test. In addition, it is a far more informative concept than the reliability coefficient in the sense that it is defined locally, as a function of ability level. It is most meaningful when the criterion test information function assumes high values for the entire range of ability in question so that the normal approximation of the conditional distribution of the maximum likelihood estimate, given ability, is readily acceptable. In computerized adaptive testing, if a constant test information is used as the criterion for terminating the presentation of new items, the standard error of estimation will be constant for the entire range of ability in question (i.e., each and every examinee's ability will be estimated with the same level of accuracy, as explained above).

The Problem of Deviated Values of the Maximum Likelihood Estimate

Throughout this conference, several researchers have voiced their concern about deviated values of the maximum likelihood estimates, and it seems fairly common that they prefer Bayesian estimates to the maximum likelihood estimate. It has been shown (Samejima, 1977b) that when the criterion test information function assumes high enough values so that the normal approximation is acceptable for the conditional distribution of the maximum likelihood estimate, given ability, the population variance of the maximum likelihood estimate is *greater* than that of the true ability for any group of examinees. To be more precise, given that

$$E(\hat{\theta}|\theta) \doteq \theta \quad , \qquad [5]$$

for any population of examinees,

$$E(\hat{\theta}) \doteq E(\theta) \qquad [6]$$

can be obtained. In addition,

$$E[\hat{\theta} - E(\hat{\theta})]^{m} \doteq \sum_{r=0}^{m} \begin{bmatrix} m \\ r \end{bmatrix} E[\{\theta - E(\theta)\}^{m-r} E\{(\hat{\theta} - \theta)^{r}|\theta\}] \qquad [7]$$

can be written for the m^{th} moment of the maximum likelihood estimate about its mean, and, in particular,

$$\text{Var.}\ (\hat{\theta}) \doteq \text{Var.}\ (\theta) + E[I(\theta)^{-1}]. \qquad [8]$$

Thus, it is obvious that the population variance of the maximum likelihood estimate is greater than that of the true ability for any group of examinees, unless the criterion test information function assumes positive infinity.

It is only natural, therefore, that more deviated values of the maximum likelihood estimate tend to be obtained, compared with true ability. Concerning this problem, the following points should be of particular consideration:

1. Although there is a tendency for researchers to be concerned over extreme values of the maximum likelihood estimate, there is no reason to say that a discrepancy of, for example, .09 from the true ability at either end of the distribution is worse than the same amount of discrepancy at the middle of the distribution.

2. This discrepancy at both ends of the distribution tends to be great, not because of the intrinsic nature of the maximum likelihood estimate, but because of low values of the test information function around these areas, which researchers unintentionally allow. This problem can easily be solved by adopting constant test information as the criterion in terminating presentation of new items in computerized adaptive testing, as was suggested above.

If these two points are kept in mind and the situation is adjusted so that the latter type of error is not allowed to happen, there will be no logical reason why deviated values of the maximum likelihood estimate should be of concern. When Bayesian estimation is used, since in many cases the prior ability distribution has a greater density around the median, the estimate for an extreme level of ability tends to regress to the median.

The question may be asked, What is the logically meaningful gain? For individual comparisons, it is only fair to give the same value of estimate to all the examinees having an identical response pattern; thus, the maximum likelihood estimate should be used. For group comparisons with some information about the individuals who belong to the groups, Bayesian estimation is more suitable, in the sense that the Bayes estimator makes the mean square error minimal.

In the latter case, however, there are in practice two difficulties. First, pre-assigning a probability density function for a group of examinees is often either impossible or extremely difficult, unless we already know about the group to be tested. In many cases, however, the examinees are tested because knowledge about these people is desired with respect to the ability to be measured, not because we already know about them. Thus, the use of a prior distribution in estimating their ability can itself be a tautology. Second, with the exception, perhaps, of young children who have not accumulated much experience in life, there is a serious difficulty in defining the population to which an individual belongs. A person who happens to be, for example, black and female with a Ph.D. in psychology could be assigned to the population of all the blacks, the population of all women, or the population of all the Ph.D.'s in psychology. In each case a different value of a Bayesian estimate would probably be obtained. Which one should be accepted in preference to the other? To avoid this, there could be an attempt to specify the population to which this individual belongs more precisely by taking the intersection of these populations. It is likely, however, that it could ultimately be said that *every adult individual belongs to his or her own unique population without sharing it with anybody else.* Thus, Bayesian estimation becomes impossible.

It is easy to use some arbitrary probability density function in Bayesian estimation in both simulation studies and actual experiments. It should be kept in mind, however, that any meaningful research should be an image of reality and that to produce an artifact in such a way is not only meaningless but harmful.

If a relatively large dispersion of the maximum likelihood estimate still produces concern, perhaps the following scale change will help. From Equation 8, the ratio of the population standard deviation of θ to that of its maximum likelihood estimate is given by

$$\{\mathrm{Var.}\ (\hat{\theta}) - E[I(\theta)^{-1}]\}^{\frac{1}{2}}\ [\mathrm{Var.}\ (\hat{\theta})]^{-\frac{1}{2}} \quad . \qquad [9]$$

Thus, by multiplying the values of the maximum likelihood estimate with this constant, the estimate will be obtained which has the same standard deviation as the true ability, and also the same origin, by virtue of Equation 5. (In practice, there exist sampling fluctuations and estimated values for $E[I(\theta)^{-1}]$ and Var. $(\hat{\theta})$ must also be used.) In computerized adaptive testing, if a constant criterion test information is used, then Equation 8 will be simplified to the form

$$\mathrm{Var.}\ (\hat{\theta}) \doteq \mathrm{Var.}\ (\theta) + \sigma^2 \quad , \qquad [10]$$

and the ratio in Equation 9 becomes

$$[\mathrm{Var.}\ (\hat{\theta}) - \sigma^2]^{\frac{1}{2}}\ [\mathrm{Var.}\ (\hat{\theta})]^{-\frac{1}{2}} \quad , \qquad [11]$$

which is even easier to handle.

Use can also be made of some other estimators not requiring a prior distribution of ability and yet having dispersions which tend to be smaller than that of the maximum likelihood estimate. Such an estimator has been hinted previously (Samejima, 1977c); it is derived from the conditional moments of the true ability, given its maximum likelihood estimate. A detailed discussion will be made elsewhere, however.

Challenging the Multiple-Choice Item

The problem of dealing with the multiple-choice test item is not simple. It is handled by researchers fairly casually, however, by adopting the three-parameter normal or logistic model (Birnbaum, 1968) without questioning its adequacy. Very few of them really attempt to go back to the model's basic principle or question whether or not it is legitimate to use the model for their particular multiple-choice data.

It has been pointed out (Samejima, 1972, 1973) that even with the three-parameter normal ogive or logistic model, there are some theoretical difficulties caused by random guessing. What should be emphasized, however, is the fact that the model is so carelessly used by researchers.

The three-parameter normal or logistic model is based on the knowledge, or random guessing, principle (i.e., the principle which states that if the examinee knows the answer to the question, he/she will give the correct answer; and if not, guess randomly). Following this principle in the multiple-choice situation, the third parameter (i.e., the lower asymptote of the item characteristic function) should be unity divided by the number of alternatives given in the item. Although the model is built on this principle, many researchers still accept it, even if they obtain a substantially different value (usually lower) for the third parameter as the result of analysis of their multiple-choice data. This is obviously contradictory.

Another easy way to find out about the fit of the model to the data is to study the frequency distribution of the wrong answers. Since the knowledge or random guessing principle does not include any effects of the wrong answers which are used as alternatives in addition to the right answer, a uniform distribution should be expected for these wrong answers. A simple chi-square test will examine it easily. In my experience, the null hypothesis of the uniform distribution has to be rejected in many cases, and one or more alternatives is found having a far greater frequency than the others. This is quite natural, since in test construction distractors (i.e., incorrect but plausible answers) are included among the alternatives so that the correct answer is not made too conspicuous. These two facts indicate that the examinee's behavior in the multiple-choice situation is affected by distractors and that the knowledge or random guessing principle is not working; therefore, the adoption of the three-parameter normal or logistic model would not be legitimate.

This is a very serious problem. Although some results can always be produced even if wrong models are adopted, such results are nothing but artifacts. Conscientious researchers should always be careful to check the fit of the model by all available means before they decide to adopt it.

At this stage, what is needed most of all is psychometricians' efforts to investigate the examinee's behavior in the multiple-choice situation. This would eventually lead to the construction of new models for the multiple-choice item. For some reason, this has not been done sufficiently yet; and researchers tend to use the three-parameter models without questioning their legitimacy.

The knowledge and random guessing principle and the three-parameter model could be considered as one extreme, and the individual choice behavior and Bock's model (Bock, 1972) could be considered as the other extreme. It is conceivable that the pursuit of the examinee's behavior in the multiple-choice situation will end up with the construction of several new models which lie between these two extremes or that completely new models may be discovered. In any case, at this stage such a challenge is badly needed; otherwise, researchers may spend their time producing nothing worthwhile.

Multidimensional Latent Trait Theory

Throughout this conference, some researchers have shown their strong interest in multidimensional latent trait theory. This is very natural, since the theory will find many uses in computerized adaptive testing in the future. Unfortunately, the mathematics involved in multidimensional latent trait theory are much more complicated than in unidimensional latent trait theory. Consequently, the computer programming for this purpose becomes very difficult, or almost impossible, unless it follows some continuous response model in which simple sufficient statistics exist (Samejima, 1974). Generally speaking, as the dimensionality increases, not only does the programming become increasingly complicated but also a substantial increase in computer time occurs.

In spite of these difficulties, multidimensional latent trait theory needs to be developed further so that good use of it can be made in computerized adaptive testing. At this stage, theorists must be depended upon to do more work.

References

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

Bock, R. D. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 1972, 37, 29-51.

Samejima, F. A general model for free-response data. Psychometrika Monograph, No. 18, 1972.

Samejima, F. A comment on Birnbaum's three-parameter logistic model in the latent trait theory. Psychometrika, 1973, 38, 221-233.

Samejima, F. Normal ogive model on the continuous response level in the multidimensional latent space. Psychometrika, 1974, 39, 111-121.

Samejima, F. A use of the information function in tailored testing. Applied Psychological Measurement, 1977, 1, 233-247. (a)

Samejima, F. A method of estimating item characteristic functions using the maximum likelihood estimate of ability. Psychometrika, 1977, 42, 163-191. (b)

Samejima, F. Comparison of two methods of estimating operating characteristics using the maximum likelihood estimate of latent trait. Paper presented at Psychometric Society Spring Meeting, Chapel Hill, 1977. (c)

VERN URRY
U.S. CIVIL SERVICE COMMISSION

Two years ago the Conference on Computerized Adaptive Testing was held in Washington, DC, during which time a dialogue was opened, concentrating on ability testing. At the time of the present conference, the major psychometric problems of ability testing have been solved. An interdisciplinary dialogue has been opened with test administrators, systems analysts, and budget managers; and major interdisciplinary problems are being isolated, which is a first and necessary step to their solution. The scope of the conference has been expanded to other fruitful areas of psychological measurement. These past two years thus have been marked by significant progress.

Crucial to further development or implementaion of ability testing is an improved interdisciplinary dialogue; this is necessary if research results are to be put into operation. Other developments that are important are interfaces between tailored ability testing and operational selection, placement, classification, and career counseling. Data accumulated in selection, placement and classification can be used to assist in career counseling. There would be, in conjunction, job availability information which would make this more meaningful.

The approach to selection, placement, and classification should employ utility theory, as introduced by Hubert Brogden (1949) in an article entitled "When Testing Pays Off." Brogden's developments were later given greater sophistication by Cronbach and Gleser (1965) in Psychological Tests and Personnel Decisions. Use of this approach will improve our ability to communicate with those who will make the financial decisions regarding tailored testing, since the benefits as well as costs can be communicated in terms of dollars.

If one is to make a prediction, it must be of a future both exciting and challenging. Given the current rate of achievement, the next decade will be marked with outstanding progress for psychological measurement.

References

Brogden, H. E. When testing pays off. Personnel Psychology, 1949, 2, 171-183.

Cronbach, L. J., & Gleser, G. C. Psycholgoical tests and personnel decisions. Urbana, IL: University of Illinois Press, 1965.

DAVID J. WEISS
UNIVERSITY OF MINNESOTA

Following Lord's approach, I am not going to speculate on the future of adaptive testing. Instead, I am going to tell you what will happen in the future; specifically, the future direction of our research program. We are presently doing research on five areas of ability testing. The first area is the investigation of unidimensional branching strategies. We are coming very close to answers to some basic questions about what is and what is not a good branching strategy. Already, many of the strategies that I identified in my reveiws (Weiss, 1973, 1974) are no longer being actively studied. And some of the strategies currently being used will also fall by the wayside in the next several years. At the same time, new unidimensional strategies will be developed that have not as yet been conceived, so that in the near future we will know the best way to unidimensionally branch through an item pool; then that problem will be forgotten until such a time when new strategies are devised that perform better than even the best of what is presently known.

The second concern of our research program on adaptive ability testing is that of psychological effects. In the future, not only will tests be adapted to each individual's ability level, achievement level, or personality level as testing is in progress but also tests will be adapted to individual differences in other kinds of characteristics. In addition, an optimal set of characteristics for each individual will be identified that will provide not only the best psychometric estimate of an individual's scores but also the best psychological estimate (to the extent that the individual feels good on completing the test). Whether it be feedback, difficulty level, lowering anxiety, raising motivation, or a host of other variables that have not as yet been identified, tests will be adapted to individual differences in a number of domains at the same time as measurement accuracy is increased through the identification of psychometrically optimal branching strategies.

Third, the multiple-choice item will disappear and exist only in museums. We will learn how to use graded responses, continuous responses, and free responses; and in the process we will humanize testing even a little bit more, not only by adapting the test to individual differences in abilities and other variables, but also by allowing people to respond in a more natural way than is allowed by multiple-choice tests.

Our fourth area of research is multidimensionality, in which the problems of multidimensional branching will be solved. Initially, this will be approached by branching between subtests; then, the problem will be approached by multidimensionally branching an individual in a multivariate space simultaneously. The problem will be resolved by integration of these approaches with the approach that Sympson has described in which test items are treated as being multidimensional. All of these things will converge in some number of years, and we will utilize the true multidimensional nature of our ability test items and our ability batteries in a most efficient way that can not now be conceptualized.

Finally, in the ability testing area, computerized adaptive testing will become the mechanism by which testing becomes integrated into the mainstream of psychology. Psychometricians and cognitive information processing researchers are going to begin interacting with each other. The seeds of this can be seen now in the work of John Carroll at the University of North Carolina and Hunt and Lunneborg at the University of Washington. It is also evident in our work here at the University of Minnesota in which some of the data obtained from the computerized administration of tests (e.g., the latencies) will begin to be used in the same kinds of ways that information processing researchers will use them. The difference now, however, is that we are concerned with individual differences and they are not. Cognitive psychologists will begin to be concerned with individual differences,and psychometricians will begin to use some of their methodologies. This marriage is being mediated by the computer, which cognitive psychologists have been using for some time in various ways to measure information processing abilities and other aspects of cognitive development that,through computerized adaptive testing,are now becoming of interest to psychometricians.

In the area of achievement testing, we will be measuring,not by criterion-referenced tests, not by norm-referenced tests, not by time-referenced tests,but by considering the individual himself/herself as the point of reference. We will use latent trait theory and computerized administration to measure whether or not an individual has learned anything, which is the real objective of achievement testing. Latent trait theory combined with computerized adaptive testing techniques will be used to describe individual learning curves.

In the area of achievement testing within computer-assisted instruction (which is the direction toward which instruction will slowly evolve), latent trait theory will be used to incorporate for measurement purposes all the information that a person provides in the process of instruction. Thus, every response a student makes on every frame in a CAI lesson will be treated as measurement information by applying multidimensional models. Each response will be treated as a multidimensional response;and that information will be used to measure individual differences,using data obtained as an integral part of instruction. Thus,testing and instruction will merge and will no longer be treated as separate activities.

These are some of the current objectives for our adaptive testing research program at the Universtiy of Minnesota. We expect to make substantial progress in all of these areas in the near future in a concentrated effort to make computerized adaptive testing of ability and achievement an important vehicle for the improvement of psychological measurement.